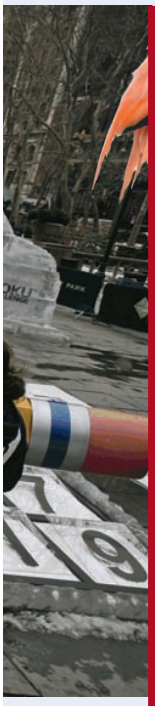
Chapter 2

Modeling Distributions of Data



case study

Do You Sudoku?

The sudoku craze has officially swept the globe. Here's what Will Shortz, crossword puzzle editor for the *New York Times*, said about sudoku:

> *As humans we seem to have an innate desire to fill up empty spaces. This might explain part of the appeal of sudoku, the new international craze, with its empty squares to be filled with digits. Since April 2005, when sudoku was introduced to the United States in* The New York Post, *more than half the leading American newspapers have begun printing one or more sudoku a day. No puzzle has had such a fast introduction in newspapers since the crossword craze of 1924–25.*[1]

Since then, millions of people have made sudoku part of their daily routines.

One of the authors played an online game of sudoku at www.websudoku.com. The graph provides information about how well he did. (His time is marked with an arrow.)

Easy level average time: 5 minutes, 6 seconds.

In this chapter, you'll learn more about how to describe the location of an individual observation—like the author's sudoku time—within a distribution.

Introduction

Suppose Jenny earns an 86 (out of 100) on her next statistics test. Should she be satisfied or disappointed with her performance? That depends on how her score compares with the scores of the other students who took the test. If 86 is the highest score, Jenny might be very pleased. Maybe her teacher will "curve" the grades so that Jenny's 86 becomes an "A." But if Jenny's 86 falls below the "average" in the class, she may not be so happy.

Section 2.1 focuses on describing the location of an individual within a distribution. We begin by discussing a familiar measure of position: *percentiles*. Next, we introduce a new type of graph that is useful for displaying percentiles. Then we consider another way to describe an individual's position that is based on the mean and standard deviation. In the process, we examine the effects of transforming data on the shape, center, and spread of a distribution.

Sometimes it is helpful to use graphical models called *density curves* to describe the location of individuals within a distribution, rather than relying on actual data values. Such models are especially helpful when data fall in a bell-shaped pattern called a *Normal distribution*. Section 2.2 examines the properties of Normal distributions and shows you how to perform useful calculations with them.

ACTIVITY | Where do I stand?

MATERIALS:
Masking tape to mark number line scale

In this Activity, you and your classmates will explore ways to describe where you stand (literally!) within a distribution.

1. Your teacher will mark out a number line on the floor with a scale running from about 58 to 78 inches.
2. Make a human dotplot. Each member of the class should stand at the appropriate location along the number line scale based on height (to the nearest inch).
3. Your teacher will make a copy of the dotplot on the board for your reference.
4. What percent of the students in the class have heights less than yours? This is your *percentile* in the distribution of heights.

5. Work with a partner to calculate the mean and standard deviation of the class's height distribution from the dotplot. Confirm these values with your classmates.
6. Where does your height fall in relation to the mean: above or below? How far above or below the mean is it? How many standard deviations above or below the mean is it? This last number is the *z-score* corresponding to your height.
7. *Class discussion:* What would happen to the class's height distribution if you converted each data value from inches to centimeters? (There are 2.54 centimeters in 1 inch.) How would this change of units affect the measures of center, spread, and location (percentile and *z*-score) that you calculated?

Want to know more about where you stand—in terms of height, weight, or even body mass index? Do a Web search for "Clinical Growth Charts" at the National Center for Health Statistics site, www.cdc.gov/nchs.

2.1 Describing Location in a Distribution

WHAT YOU WILL LEARN **By the end of the section, you should be able to:**

- Find and interpret the percentile of an individual value within a distribution of data.
- Estimate percentiles and individual values using a cumulative relative frequency graph.
- Find and interpret the standardized score (*z*-score) of an individual value within a distribution of data.
- Describe the effect of adding, subtracting, multiplying by, or dividing by a constant on the shape, center, and spread of a distribution of data.

Here are the scores of all 25 students in Mr. Pryor's statistics class on their first test:

79 81 80 77 73 83 74 93 78 80 75 67 73
77 83 **86** 90 79 85 83 89 84 82 77 72

The bold score is Jenny's **86**. How did she perform on this test relative to her classmates?

The stemplot displays this distribution of test scores. Notice that the distribution is roughly symmetric with no apparent outliers. From the stemplot, we can see that Jenny did better than all but three students in the class.

```
6 | 7
7 | 2334
7 | 5777899
8 | 00123334
8 | 569
9 | 03
```

Key: 7|2 is a student who scored 72 on the test

Measuring Position: Percentiles

One way to describe Jenny's location in the distribution of test scores is to tell what percent of students in the class earned scores that were below Jenny's score. That is, we can calculate Jenny's **percentile**.

DEFINITION: Percentile

The ***p*th percentile** of a distribution is the value with *p* percent of the observations less than it.

Using the stemplot, we see that Jenny's 86 places her fourth from the top of the class. Because 21 of the 25 observations (84%) are below her score, Jenny is at the 84th percentile in the class's test score distribution.

EXAMPLE Mr. Pryor's First Test

Finding percentiles

PROBLEM: Use the scores on Mr. Pryor's first statistics test to find the percentiles for the following students:

(a) Norman, who earned a 72.

(b) Katie, who scored 93.

(c) The two students who earned scores of 80.

SOLUTION:

(a) Only 1 of the 25 scores in the class is below Norman's 72. His percentile is computed as follows: 1/25 = 0.04, or 4%. So Norman scored at the 4th percentile on this test.

(b) Katie's 93 puts her at the 96th percentile, because 24 out of 25 test scores fall below her result.

(c) Two students scored an 80 on Mr. Pryor's first test. Because 12 of the 25 scores in the class were less than 80, these two students are at the 48th percentile.

For Practice *Try Exercise* 1

Note: Some people define the pth percentile of a distribution as the value with p percent of observations *less than or equal to* it. Using this alternative definition of percentile, it is possible for an individual to fall at the 100th percentile. If we used this definition, the two students in part (c) of the example would fall at the 56th percentile (14 of 25 scores were less than or equal to 80). Of course, because 80 is the median score, it is also possible to think of it as being the 50th percentile. Calculating percentiles is not an exact science, especially with small data sets! We'll stick with the definition of percentile we gave earlier for consistency.

Cumulative Relative Frequency Graphs

Age	Frequency
40–44	2
45–49	7
50–54	13
55–59	12
60–64	7
65–69	3

There are some interesting graphs that can be made with percentiles. One of the most common graphs starts with a frequency table for a quantitative variable. For instance, the frequency table in the margin summarizes the ages of the first 44 U.S. presidents when they took office.

Let's expand this table to include columns for relative frequency, cumulative frequency, and cumulative relative frequency.

- To get the values in the *relative frequency* column, divide the count in each class by 44, the total number of presidents. Multiply by 100 to convert to a percent.
- To fill in the *cumulative frequency* column, add the counts in the frequency column for the current class and all classes with smaller values of the variable.
- For the *cumulative relative frequency* column, divide the entries in the cumulative frequency column by 44, the total number of individuals. Multiply by 100 to convert to a percent.

Here is the original frequency table with the relative frequency, cumulative frequency, and cumulative relative frequency columns added.

Age	Frequency	Relative frequency	Cumulative frequency	Cumulative relative frequency
40–44	2	2/44 = 0.045, or 4.5%	2	2/44 = 0.045, or 4.5%
45–49	7	7/44 = 0.159, or 15.9%	9	9/44 = 0.205, or 20.5%
50–54	13	13/44 = 0.295, or 29.5%	22	22/44 = 0.500, or 50.0%
55–59	12	12/44 = 0.273, or 27.3%	34	34/44 = 0.773, or 77.3%
60–64	7	7/44 = 0.159, or 15.9%	41	41/44 = 0.932, or 93.2%
65–69	3	3/44 = 0.068, or 6.8%	44	44/44 = 1.000, or 100%

Some people refer to cumulative relative frequency graphs as "ogives" (pronounced "o-jives").

To make a **cumulative relative frequency graph,** we plot a point corresponding to the cumulative relative frequency in each class at the smallest value of the *next* class. For example, for the 40 to 44 class, we plot a point at a height of 4.5% above the age value of 45. This means that 4.5% of presidents were inaugurated *before* they were 45 years old. (In other words, age 45 is the 4.5th percentile of the inauguration age distribution.)

It is customary to start a cumulative relative frequency graph with a point at a height of 0% at the smallest value of the first class (in this case, 40). The last point we plot should be at a height of 100%. We connect consecutive points with a line segment to form the graph. Figure 2.1 shows the completed cumulative relative frequency graph.

Here's an example that shows how to interpret a cumulative relative frequency graph.

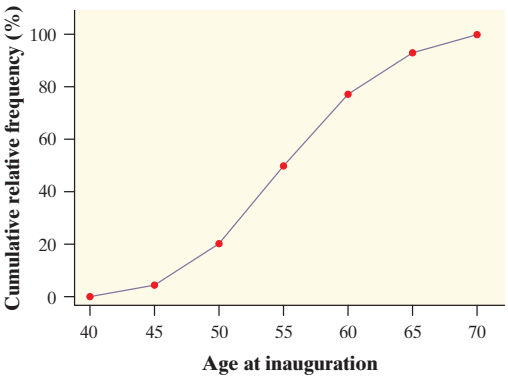


FIGURE 2.1 Cumulative relative frequency graph for the ages of U.S. presidents at inauguration.

EXAMPLE Age at Inauguration

Interpreting a cumulative relative frequency graph

What can we learn from Figure 2.1? The graph grows very gradually at first because few presidents were inaugurated when they were in their 40s. Then the graph gets very steep beginning at age 50. Why? Because most U.S. presidents were in their 50s when they were inaugurated. The rapid growth in the graph slows at age 60.

Suppose we had started with only the graph in Figure 2.1, without any of the information in our original frequency table. Could we figure out what percent of presidents were between 55 and 59 years old at their inaugurations? Sure. Because the point at age 60 has a cumulative relative frequency of about 77%, we know that about 77% of presidents were inaugurated before they were 60 years old. Similarly, the point at age 55 tells us that about 50% of presidents were younger than 55 at inauguration. As a result, we'd estimate that about 77% − 50% = 27% of U.S. presidents were between 55 and 59 when they were inaugurated.

A cumulative relative frequency graph can be used to describe the position of an individual within a distribution or to locate a specified percentile of the distribution, as the following example illustrates.

EXAMPLE Ages of U.S. Presidents

Interpreting cumulative relative frequency graphs

PROBLEM: Use the graph in Figure 2.1 on the previous page to help you answer each question.

(a) Was Barack Obama, who was first inaugurated at age 47, unusually young?

(b) Estimate and interpret the 65th percentile of the distribution.

SOLUTION:

(a) To find President Obama's location in the distribution, we draw a vertical line up from his age (47) on the horizontal axis until it meets the graphed line. Then we draw a horizontal line from this point of intersection to the vertical axis. Based on Figure 2.2(a), we would estimate that Barack Obama's inauguration age places him at the 11% cumulative relative frequency mark. That is, he's at the 11th percentile of the distribution. In other words, about 11% of all U.S. presidents were younger than Barack Obama when they were inaugurated and about 89% were older.

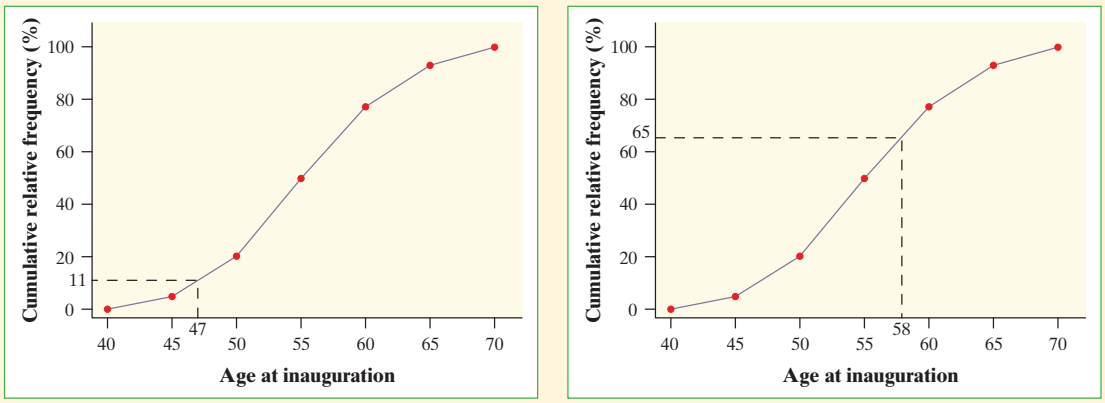


FIGURE 2.2 The cumulative relative frequency graph of presidents' ages at inauguration is used to (a) locate President Obama within the distribution and (b) determine the 65th percentile, which is about 58 years.

(b) The 65th percentile of the distribution is the age with cumulative relative frequency 65%. To find this value, draw a horizontal line across from the vertical axis at a height of 65% until it meets the graphed line. From the point of intersection, draw a vertical line down to the horizontal axis. In Figure 2.2(b), the value on the horizontal axis is about 58. So about 65% of all U.S. presidents were younger than 58 when they took office.

For Practice *Try Exercise* 9

THINK ABOUT IT

Percentiles and quartiles: Have you made the connection between percentiles and the quartiles from Chapter 1? Earlier, we noted that the median (second quartile) corresponds to the 50th percentile. What about the first quartile, Q_1? It's at the median of the lower half of the ordered data, which puts it about one-fourth of the way through the distribution. In other words, Q_1 is roughly the 25th percentile. By similar reasoning, Q_3 is approximately the 75th percentile of the distribution.

CHECK YOUR UNDERSTANDING

1. *Multiple choice: Select the best answer.* Mark receives a score report detailing his performance on a statewide test. On the math section, Mark earned a raw score of 39, which placed him at the 68th percentile. This means that

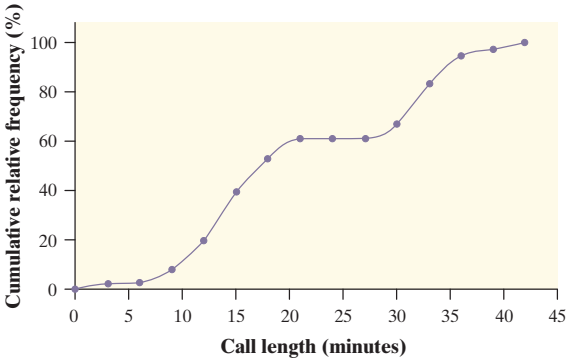
(a) Mark did better than about 39% of the students who took the test.

(b) Mark did worse than about 39% of the students who took the test.

(c) Mark did better than about 68% of the students who took the test.

(d) Mark did worse than about 68% of the students who took the test.

(e) Mark got fewer than half of the questions correct on this test.



2. Mrs. Munson is concerned about how her daughter's height and weight compare with those of other girls of the same age. She uses an online calculator to determine that her daughter is at the 87th percentile for weight and the 67th percentile for height. Explain to Mrs. Munson what this means.

Questions 3 and 4 relate to the following setting. The graph displays the cumulative relative frequency of the lengths of phone calls made from the mathematics department office at Gabalot High last month.

3. About what percent of calls lasted less than 30 minutes? 30 minutes or more?

4. Estimate Q_1, Q_3, and the IQR of the distribution.

Measuring Position: *z*-Scores

6	7
7	2334
7	5777899
8	00123334
8	569
9	03

Key: 7|2 is a student who scored 72 on the test

Let's return to the data from Mr. Pryor's first statistics test, which are shown in the stemplot. Figure 2.3 provides numerical summaries from Minitab for these data. Where does Jenny's score of 86 fall relative to the mean of this distribution? Because the mean score for the class is 80, we can see that Jenny's score is "above average." But how much above average is it?

We can describe Jenny's location in the distribution of her class's test scores by telling how many standard deviations above or below the mean her score is. Because the mean is 80 and the standard deviation is about 6, Jenny's score of 86 is about one standard deviation above the mean.

Converting observations like this from original values to standard deviation units is known as **standardizing**. To standardize a value, subtract the mean of the distribution and then divide the difference by the standard deviation.

The relationship between the mean and the median is about what you'd expect in this fairly symmetric distribution.

Minitab

Session

Descriptive Statistics: Test 1 scores

Variable	N	Mean	Median	TrMean	StDev	SE Mean
Test 1 scores	25	80.00	80.00	80.00	6.07	1.21

Variable	Minimum	Maximum	Q1	Q3
Test 1 scores	67.00	93.00	76.00	83.50

FIGURE 2.3 Minitab output for the scores of Mr. Pryor's students on their first statistics test.

DEFINITION: Standardized score (*z*-score)

If x is an observation from a distribution that has known mean and standard deviation, the **standardized score** for x is

$$z = \frac{x - \text{mean}}{\text{standard deviation}}$$

A standardized score is often called a ***z*-score**.

A z-score tells us how many standard deviations from the mean an observation falls, and in what direction. Observations larger than the mean have positive z-scores. Observations smaller than the mean have negative z-scores. For example, Jenny's score on the test was $x = 86$. Her *standardized* score (z-score) is

$$z = \frac{x - \text{mean}}{\text{standard deviation}} = \frac{86 - 80}{6.07} = 0.99$$

That is, Jenny's test score is 0.99 standard deviations *above* the mean score of the class.

EXAMPLE Mr. Pryor's First Test, Again

Finding and interpreting z-scores

PROBLEM: Use Figure 2.3 on the previous page to find the standardized scores (z-scores) for each of the following students in Mr. Pryor's class. Interpret each value in context.

(a) Katie, who scored 93.

(b) Norman, who earned a 72.

SOLUTION:

(a) Katie's 93 was the highest score in the class. Her corresponding z-score is

$$z = \frac{93 - 80}{6.07} = 2.14$$

In other words, Katie's result is 2.14 standard deviations above the mean score for this test.

(b) For Norman's 72, his standardized score is

$$z = \frac{72 - 80}{6.07} = -1.32$$

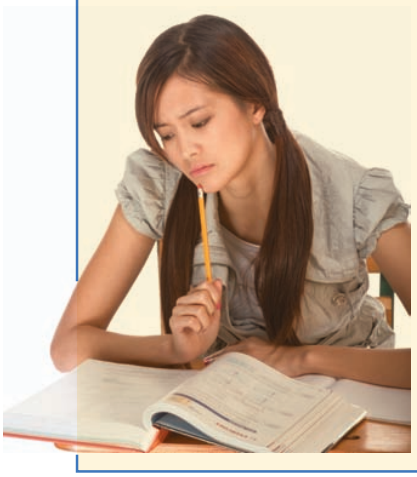
Norman's score is 1.32 standard deviations below the class mean of 80.

For Practice *Try Exercise* 15(b)

We can also use z-scores to compare the position of individuals in different distributions, as the following example illustrates.

EXAMPLE Jenny Takes Another Test

Using z-scores for comparisons



The day after receiving her statistics test result of 86 from Mr. Pryor, Jenny earned an 82 on Mr. Goldstone's chemistry test. At first, she was disappointed. Then Mr. Goldstone told the class that the distribution of scores was fairly symmetric with a mean of 76 and a standard deviation of 4.

PROBLEM: On which test did Jenny perform better relative to the class? Justify your answer.

SOLUTION: Jenny's *z*-score for her chemistry test result is

$$z = \frac{82 - 76}{4} = 1.50$$

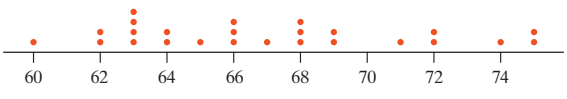
Her 82 in chemistry was 1.5 standard deviations above the mean score for the class. Because she scored only 0.99 standard deviations above the mean on the statistics test, Jenny did better relative to the class in chemistry.

For Practice *Try Exercise* 11

We often standardize observations to express them on a common scale. We might, for example, compare the heights of two children of different ages by calculating their *z*-scores. At age 2, Jordan is 89 centimeters (cm) tall. Her height puts her at a *z*-score of 0.5; that is, she is one-half standard deviation above the mean height of 2-year-old girls. Zayne's height at age 3 is 101 cm, which yields a *z*-score of 1. In other words, he is one standard deviation above the mean height of 3-year-old boys. So Zayne is taller relative to boys his age than Jordan is relative to girls her age. The standardized heights tell us where each child stands (pun intended!) in the distribution for his or her age group.

CHECK YOUR UNDERSTANDING

Mrs. Navard's statistics class has just completed the first three steps of the "Where Do I Stand?" Activity (page 84). The figure below shows a dotplot of the class's height distribution, along with summary statistics from computer output.



Variable	n	$\bar{x}$	s_x	Min	Q_1	Med	Q_3	Max
Height	25	67	4.29	60	63	66	69	75

1. Lynette, a student in the class, is 65 inches tall. Find and interpret her *z*-score.
2. Another student in the class, Brent, is 74 inches tall. How tall is Brent compared with the rest of the class? Give appropriate numerical evidence to support your answer.
3. Brent is a member of the school's basketball team. The mean height of the players on the team is 76 inches. Brent's height translates to a *z*-score of −0.85 in the team's height distribution. What is the standard deviation of the team members' heights?

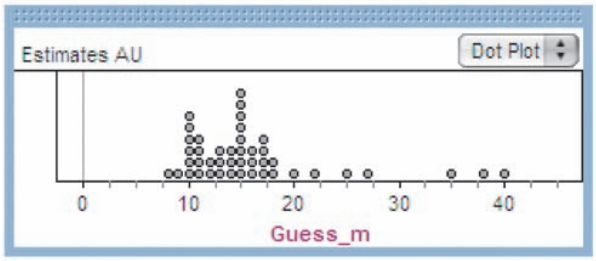
Transforming Data

To find the standardized score (z-score) for an individual observation, we transform this data value by subtracting the mean and dividing the difference by the standard deviation. Transforming converts the observation from the original units of measurement (inches, for example) to a standardized scale. What effect do these kinds of transformations—adding or subtracting; multiplying or dividing—have on the shape, center, and spread of the entire distribution? Let's investigate using an interesting data set from "down under."

Soon after the metric system was introduced in Australia, a group of students was asked to guess the width of their classroom to the nearest meter. Here are their guesses in order from lowest to highest:[2]

8	9	10	10	10	10	10	10	11	11	11	11	12
12	13	13	13	14	14	14	15	15	15	15	15	15
15	15	16	16	16	17	17	17	17	18	18	20	22
25	27	35	38	40								

Figure 2.4 includes a dotplot of the data and some numerical summaries.



	n	$\bar{x}$	s_x	Min	Q_1	Med	Q_3	Max	IQR	Range
Guess	44	16.02	7.14	8	11	15	17	40	6	32

FIGURE 2.4 Fathom dotplot and summary statistics for Australian students' guesses of the classroom width.

Let's practice what we learned in Chapter 1 and describe what we see.

Shape: The distribution of guesses appears skewed to the right and bimodal, with peaks at 10 and 15 meters.

Center: The median guess was 15 meters and the mean guess was about 16 meters. Due to the clear skewness and potential outliers, the median is a better choice for summarizing the "typical" guess.

Spread: Because $Q_1 = 11$, about 25% of the students estimated the width of the room to be fewer than 11 meters. The 75th percentile of the distribution is at about $Q_3 = 17$. The IQR of 6 meters describes the spread of the middle 50% of students' guesses. The standard deviation tells us that the typical distance of students' guesses from the mean was about 7 meters. Because s_x is not resistant to extreme values, we prefer the IQR to describe the variability of this distribution.

Outliers: By the $1.5 \times IQR$ rule, values greater than $17 + 9 = 26$ meters or less than $11 - 9 = 2$ meters are identified as outliers. So the four highest guesses—which are 27, 35, 38, and 40 meters—are outliers.

Effect of adding or subtracting a constant: By now, you're probably wondering what the actual width of the room was. In fact, it was 13 meters wide. How close were students' guesses? The student who guessed 8 meters was too low by 5 meters. The student who guessed 40 meters was too high by 27 meters (and probably needs to study the metric system more carefully). We can examine the distribution of students' guessing errors by defining a new variable as follows:

$$\text{error} = \text{guess} - 13$$

That is, we'll subtract 13 from each observation in the data set. Try to predict what the shape, center, and spread of this new distribution will be. Refer to Figure 2.4 as needed.

EXAMPLE Estimating Room Width

Effect of subtracting a constant

Let's see how accurate your predictions were (you did make predictions, right?). Figure 2.5 shows dotplots of students' original guesses and their errors on the same scale. We can see that the original distribution of guesses has been shifted to the left. By how much? Because the peak at 15 meters in the original graph is located at 2 meters in the error distribution, the original data values have been translated 13 units to the left. That should make sense: we calculated the errors by subtracting the actual room width, 13 meters, from each student's guess.

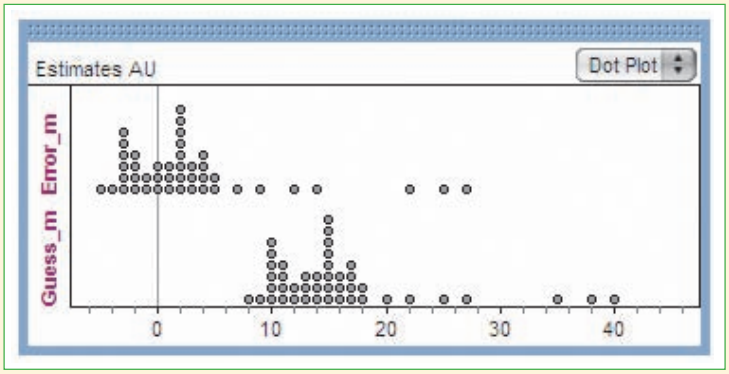


FIGURE 2.5 Fathom dotplots of students' original guesses of classroom width and the errors in their guesses.

From Figure 2.5, it seems clear that subtracting 13 from each observation did not affect the shape or spread of the distribution. But this transformation appears to have decreased the center of the distribution by 13 meters. The summary statistics in the table below confirm our beliefs.

	n	$\bar{x}$	s_x	Min	Q_1	Med	Q_3	Max	IQR	Range
Guess (m)	44	16.02	7.14	8	11	15	17	40	6	32
Error (m)	44	3.02	7.14	-5	-2	2	4	27	6	32

The error distribution is centered at a value that is clearly positive—the median error is 2 meters and the mean error is about 3 meters. So the students generally tended to overestimate the width of the room.

As the example shows, subtracting the same positive number from each value in a data set shifts the distribution to the left by that number. Adding a positive constant to each data value would shift the distribution to the right by that constant.

Let's summarize what we've learned so far about transforming data.

EFFECT OF ADDING (OR SUBTRACTING) A CONSTANT

Adding the same positive number a to (subtracting a from) each observation

- adds a to (subtracts a from) measures of center and location (mean, median, quartiles, percentiles), but
- does not change the shape of the distribution or measures of spread (range, *IQR*, standard deviation).

Effect of multiplying or dividing by a constant: Because our group of Australian students is having some difficulty with the metric system, it may not be helpful to tell them that their guesses tended to be about 2 to 3 meters too high. Let's convert the error data to feet before we report back to them. There are roughly 3.28 feet in a meter. So for the student whose error was −5 meters, that translates to

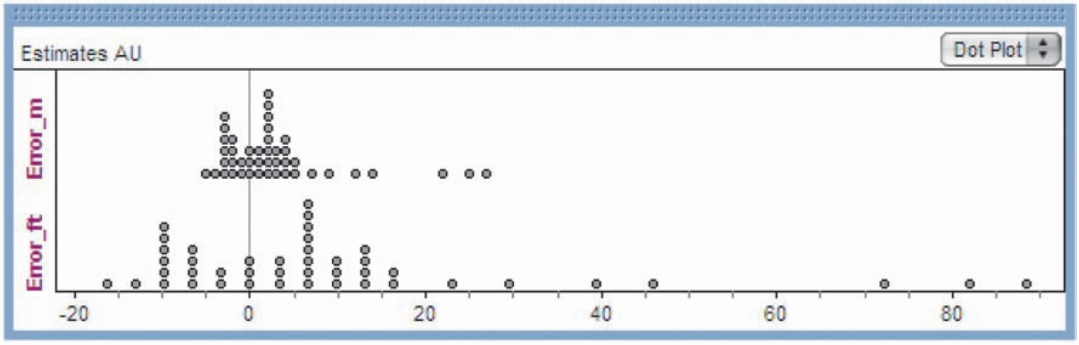
$$-5 \text{ meters} \times \frac{3.28 \text{ feet}}{1 \text{ meter}} = -16.4 \text{ feet}$$

To change the units of measurement from meters to feet, we multiply each of the error values by 3.28. What effect will this have on the shape, center, and spread of the distribution? (Go ahead, make some predictions!)

EXAMPLE Estimating Room Width

Effect of multiplying by a constant

Figure 2.6 includes dotplots of the students' guessing errors in meters and feet, along with summary statistics from computer software. The shape of the two distributions is the same—right-skewed and bimodal. However, the centers and spreads



	n	$\bar{x}$	s_x	Min	Q_1	Med	Q_3	Max	*IQR*	Range
Error (m)	44	3.02	7.14	−5	−2	2	4	27	6	32
Error (ft)	44	9.91	23.43	−16.4	−6.56	6.56	13.12	88.56	19.68	104.96

FIGURE 2.6 Fathom dotplots and numerical summaries of students' errors guessing the width of their classroom in meters and feet.

of the two distributions are quite different. The bottom dotplot is centered at a value that is to the right of the top dotplot's center. Also, the bottom dotplot shows much greater spread than the top dotplot.

When the errors were measured in meters, the median was 2 and the mean was 3.02. For the transformed error data in feet, the median is 6.56 and the mean is 9.91. Can you see that the measures of center were multiplied by 3.28? That makes sense. If we multiply all the observations by 3.28, then the mean and median should also be multiplied by 3.28.

What about the spread? Multiplying each observation by 3.28 increases the variability of the distribution. By how much? You guessed it—by a factor of 3.28. The numerical summaries in Figure 2.6 show that the standard deviation, the interquartile range, and the range have been multiplied by 3.28.

We can safely tell our group of Australian students that their estimates of the classroom's width tended to be too high by about 6.5 feet. (Notice that we choose not to report the mean error, which is affected by the strong skewness and the three high outliers.)

As before, let's recap what we discovered about the effects of transforming data.

EFFECT OF MULTIPLYING (OR DIVIDING) BY A CONSTANT

Multiplying (or dividing) each observation by the same positive number b

- multiplies (divides) measures of center and location (mean, median, quartiles, percentiles) by b,
- multiplies (divides) measures of spread (range, IQR, standard deviation) by b, but
- does not change the shape of the distribution.

It is not common to multiply (or divide) each observation in a data set by a *negative* number b. Doing so would multiply (or divide) the measures of spread by the *absolute value* of b. We can't have a negative amount of variability! Multiplying or dividing by a negative number would also affect the shape of the distribution as all values would be reflected over the y axis.

Putting it all together: Adding/subtracting and multiplying/dividing: What happens if we transform a data set by both adding or subtracting a constant and multiplying or dividing by a constant? For instance, if we need to convert temperature data from Celsius to Fahrenheit, we have to use the formula °F = 9/5(°C) + 32. That is, we would multiply each of the observations by 9/5 and then add 32. As the following example shows, we just use the facts about transforming data that we've already established.

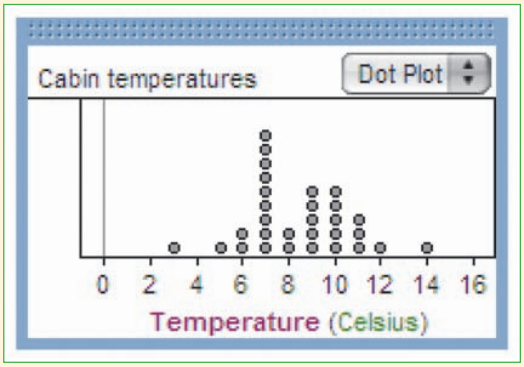
EXAMPLE

Too Cool at the Cabin?

Analyzing the effects of transformations

During the winter months, the temperatures at the Starnes's Colorado cabin can stay well below freezing (32°F or 0°C) for weeks at a time. To prevent the pipes from freezing, Mrs. Starnes sets the thermostat at 50°F. She also buys a digital thermometer that records the indoor temperature each night at midnight.

Unfortunately, the thermometer is programmed to measure the temperature in degrees Celsius. A dotplot and numerical summaries of the midnight temperature readings for a 30-day period are shown below.



	n	Mean	StDev	Min	Q_1	Median	Q_3	Max
Temperature	30	8.43	2.27	3.00	7.00	8.50	10.00	14.00

PROBLEM: Use the fact that °F = (9/5)°C + 32 to help you answer the following questions.

(a) Find the mean temperature in degrees Fahrenheit. Does the thermostat setting seem accurate?

(b) Calculate the standard deviation of the temperature readings in degrees Fahrenheit. Interpret this value in context.

(c) The 93rd percentile of the temperature readings was 12°C. What is the 93rd percentile temperature in degrees Fahrenheit?

SOLUTION:

(a) To convert the temperature measurements from Celsius to Fahrenheit, we multiply each value by 9/5 and then add 32. Multiplying the observations by 9/5 also multiplies the mean by 9/5. Adding 32 to each observation increases the mean by 32. So the mean temperature in degrees Fahrenheit is (9/5)(8.43) + 32 = 47.17°F. The thermostat doesn't seem to be very accurate. It is set at 50°F, but the mean temperature over the 30-day period is about 47°F.

(b) Multiplying each observation by 9/5 multiplies the standard deviation by 9/5. However, adding 32 to each observation doesn't affect the spread. So the standard deviation of the temperature measurements in degrees Fahrenheit is (9/5)(2.27) = 4.09°F. This means that the typical distance of the temperature readings from the mean is about 4°F. That's a lot of variation!

(c) Both multiplying by a constant and adding a constant affect the value of the 93rd percentile. To find the 93rd percentile in degrees Fahrenheit, we multiply the 93rd percentile in degrees Celsius by 9/5 and then add 32: (9/5)(12) + 32 = 53.6°F.

For Practice *Try Exercise* 19

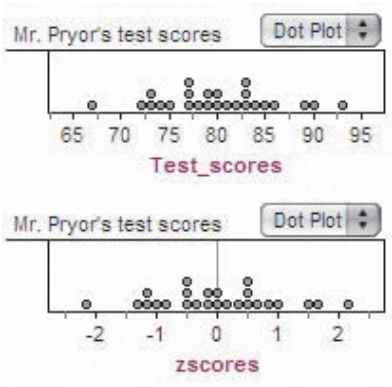
Let's look at part (c) of the example more closely. The data value of 12°C is at the 93rd percentile of the distribution, meaning that 28 of the 30 temperature readings are less than 12°C. When we transform the data, 12°C becomes 53.6°F. The value of 53.6°F is at the 93rd percentile of the transformed distribution because 28 of the 30 temperature readings are less than 53.6°F. What have we learned? Adding (or subtracting) a constant does not change an individual data

value's location within a distribution. Neither does multiplying or dividing by a positive constant.

THINK ABOUT IT

Connecting transformations and *z*-scores: What does all this transformation business have to do with *z*-scores? To standardize an observation, you subtract the mean of the distribution and then divide by the standard deviation. What if we standardized *every* observation in a distribution?

Returning to Mr. Pryor's statistics test scores, we recall that the distribution was roughly symmetric with a mean of 80 and a standard deviation of 6.07. To convert the entire class's test results to *z*-scores, we would subtract 80 from each observation and then divide by 6.07. What effect would these transformations have on the shape, center, and spread of the distribution?



This is a result worth noting! If you start with *any* set of quantitative data and convert the values to standardized scores (*z*-scores), the transformed data set will have a mean of 0 and a standard deviation of 1. The shape of the two distributions will be the same. We will use this result to our advantage in Section 2.2.

- **Shape:** The shape of the distribution of *z*-scores would be the same as the shape of the original distribution of test scores. Neither subtracting a constant nor dividing by a constant would change the shape of the graph. The dotplots confirm that the combination of these two transformations does not affect the shape.
- **Center:** Subtracting 80 from each data value would also reduce the mean by 80. Because the mean of the original distribution was 80, the mean of the transformed data would be 0. Dividing each of these new data values by 6.07 would also divide the mean by 6.07. But because the mean is now 0, dividing by 6.07 would leave the mean at 0. That is, the mean of the *z*-score distribution would be 0.
- **Spread:** The spread of the distribution would not be affected by subtracting 80 from each observation. However, dividing each data value by 6.07 would also divide our common measures of spread by 6.07. The standard deviation of the distribution of *z*-scores would therefore be 6.07/6.07 = 1.

The Minitab computer output below confirms the result: *If we standardize every observation in a distribution, the resulting set of z-scores has mean 0 and standard deviation 1.*

Descriptive Statistics: Test scores, z-scores

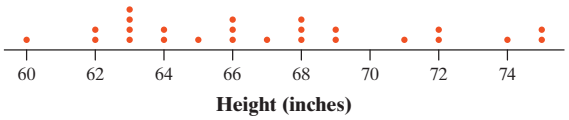
Variable	*n*	Mean	StDev	Minimum	Q_1	Median	Q_3	Maximum
Test scores	25	80.00	6.07	67.00	76.00	80.00	83.50	93.00
z-scores	25	0.00	1.00	−2.14	−0.66	0.00	0.58	2.14

Many other types of transformations can be very useful in analyzing data. We have only studied what happens when you transform data through addition, subtraction, multiplication, or division.

CHECK YOUR UNDERSTANDING

The figure on the next page shows a dotplot of the height distribution for Mrs. Navard's class, along with summary statistics from computer output.

1. Suppose that you convert the class's heights from inches to centimeters (1 inch = 2.54 cm). Describe the effect this will have on the shape, center, and spread of the distribution.



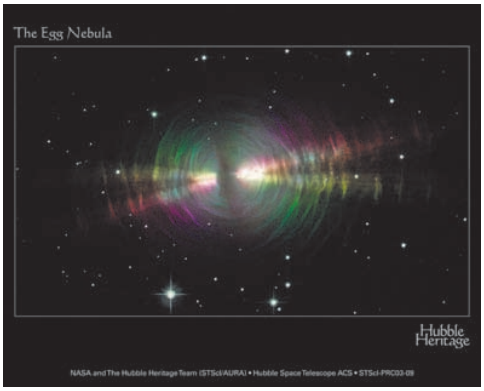
Variable	n	$\bar{x}$	s_x	Min	Q_1	Med	Q_3	Max
Height	25	67	4.29	60	63	66	69	75

2. If Mrs. Navard had the entire class stand on a 6-inch-high platform and then had the students measure the distance from the top of their heads to the ground, how would the shape, center, and spread of this distribution compare with the original height distribution?

3. Now suppose that you convert the class's heights to z-scores. What would be the shape, center, and spread of this distribution? Explain.

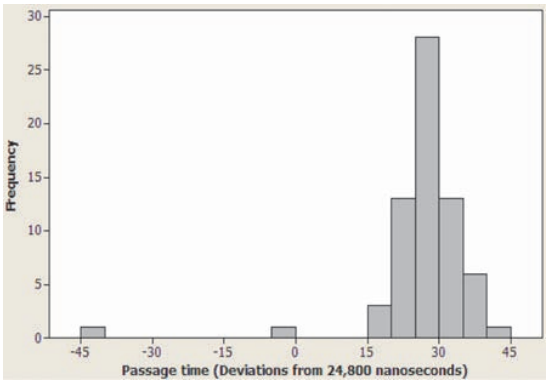
DATA EXPLORATION The speed of light

Light travels fast, but it is not transmitted instantly. Light takes over a second to reach us from the moon and over 10 billion years to reach us from the most distant objects in the universe. Because radio waves and radar also travel at the speed of light, having an accurate value for that speed is important in communicating with astronauts and orbiting satellites. An accurate value for the speed of light is also important to computer designers because electrical signals travel at light speed. The first reasonably accurate measurements of the speed of light were made more than a hundred years ago by A. A. Michelson and Simon Newcomb. The table below contains 66 measurements made by Newcomb between July and September 1882.[3]



Newcomb measured the time in seconds that a light signal took to pass from his laboratory on the Potomac River to a mirror at the base of the Washington Monument and back, a total distance of about 7400 meters. Newcomb's first measurement of the passage time of light was 0.000024828 second, or 24,828 nanoseconds. (There are 10^9 nanoseconds in a second.) The entries in the table record only the deviations from 24,800 nanoseconds.

28	26	33	24	34	−44	27	16	40	−2	29	22	24	21
25	30	23	29	31	19	24	20	36	32	36	28	25	21
28	29	37	25	28	26	30	32	36	26	30	22	36	23
27	27	28	27	31	27	26	33	26	32	32	24	39	28
24	25	32	25	29	27	28	29	16	23				



The figure provides a histogram and numerical summaries (computed with and without the two outliers) from Minitab for these data.

Descriptive Statistics: Passage time

Variable	n	Mean	Stdev	Min	Q_1	Med	Q_3	Max
P.Time	66	26.21	10.75	−44	24	27	31	40
P.Time*	64	27.75	5.08	16	24.5	27.5	31	40

1. We could convert the passage time measurements to nanoseconds by adding 24,800 to each of the data values in the table. What effect would this have on the shape, center, and spread of the distribution? Be specific.

2. After performing the transformation to nanoseconds, we could convert the measurements from nanoseconds to seconds by dividing each value by 10^9. What effect would this have on the shape, center, and spread of the distribution? Be specific.

3. Use the information provided to estimate the speed of light in meters per second. Be prepared to justify the method you used.

Section 2.1 Summary

- Two ways of describing an individual's location within a distribution are **percentiles** and ***z*-scores.** An observation's percentile is the percent of the distribution that is below the value of that observation. To standardize any observation x, subtract the mean of the distribution and then divide the difference by the standard deviation. The resulting z-score

$$z = \frac{x - \text{mean}}{\text{standard deviation}}$$

says how many standard deviations x lies above or below the distribution mean. We can also use percentiles and z-scores to compare the location of individuals in different distributions.

- A **cumulative relative frequency graph** allows us to examine location within a distribution. Cumulative relative frequency graphs begin by grouping the observations into equal-width classes (much like the process of making a histogram). The completed graph shows the accumulating percent of observations as you move through the classes in increasing order.
- It is common to **transform data**, especially when changing units of measurement. When you add a constant a to all the values in a data set, measures of center (median and mean) and location (quartiles and percentiles) increase by a. Measures of spread do not change. When you multiply all the values in a data set by a positive constant b, measures of center, location, and spread are multiplied by b. Neither of these transformations changes the shape of the distribution.

Section 2.1 Exercises

pg 86

1. **Shoes** How many pairs of shoes do students have? Do girls have more shoes than boys? Here are data from a random sample of 20 female and 20 male students at a large high school:

Female:	50	26	26	31	57	19	24	22	23	38
	13	50	13	34	23	30	49	13	15	51
Male:	14	7	6	5	12	38	8	7	10	10
	10	11	4	5	22	7	5	10	35	7

(a) Find and interpret the percentile in the female distribution for the girl with 22 pairs of shoes.

(b) Find and interpret the percentile in the male distribution for the boy with 22 pairs of shoes.

(c) Who is more unusual: the girl with 22 pairs of shoes or the boy with 22 pairs of shoes? Explain.

2. **Old folks** Here is a stemplot of the percents of residents aged 65 and older in the 50 states:

```
 7 | 0
 8 | 8
 9 | 8
10 | 019
11 | 16777
12 | 01122456778999
13 | 0001223344455689
14 | 023568
15 | 24
16 | 9
```

Key: 15|2 means 15.2% of this state's residents are 65 or older

(a) Find and interpret the percentile for Colorado, where 10.1% of the residents are aged 65 and older.

(b) Find and interpret the percentile for Rhode Island, where 13.9% of the residents are aged 65 and older.

(c) Which of these two states is more unusual? Explain.

3. **Math test** Josh just got the results of the statewide Algebra 2 test: his score is at the 60th percentile. When Josh gets home, he tells his parents that he got 60 percent of the questions correct on the state test. Explain what's wrong with Josh's interpretation.

4. **Blood pressure** Larry came home very excited after a visit to his doctor. He announced proudly to his wife, "My doctor says my blood pressure is at the 90th percentile among men like me. That means I'm better off than about 90% of similar men." How should his wife, who is a statistician, respond to Larry's statement?

5. **Growth charts** We used an online growth chart to find percentiles for the height and weight of a 16-year-old girl who is 66 inches tall and weighs 118 pounds. According to the chart, this girl is at the 48th percentile for weight and the 78th percentile for height. Explain what these values mean in plain English.

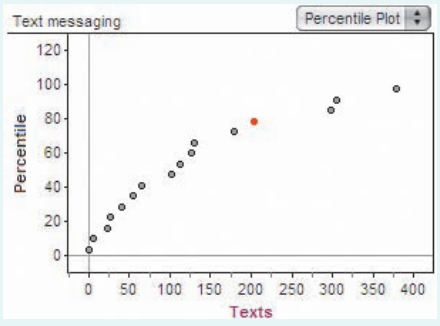
6. **Run fast** Peter is a star runner on the track team. In the league championship meet, Peter records a time that would fall at the 80th percentile of all his race times that season. But his performance places him at the 50th percentile in the league championship meet. Explain how this is possible. (Remember that lower times are better in this case!)

Exercises 7 and 8 involve a new type of graph called a ***percentile plot****. Each point gives the value of the variable being measured and the corresponding percentile for one individual in the data set.*

7. **Text me** The percentile plot below shows the distribution of text messages sent and received in a two-day period by a random sample of 16 females from a large high school.

(a) Describe the student represented by the highlighted point.

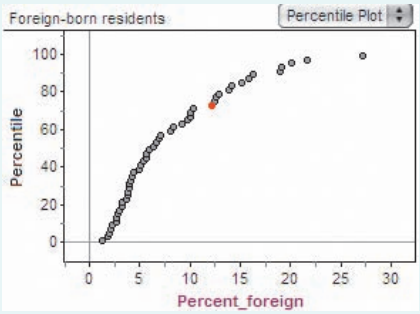
(b) Use the graph to estimate the median number of texts. Explain your method.



8. **Foreign-born residents** The following percentile plot shows the distribution of the percent of foreign-born residents in the 50 states.

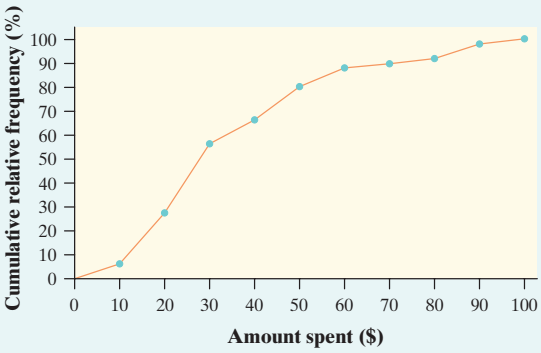
(a) The highlighted point is for Maryland. Describe what the graph tells you about this state.

(b) Use the graph to estimate the 30th percentile of the distribution. Explain your method.



9. **Shopping spree** The figure below is a cumulative relative frequency graph of the amount spent by 50 consecutive grocery shoppers at a store.

pg 88

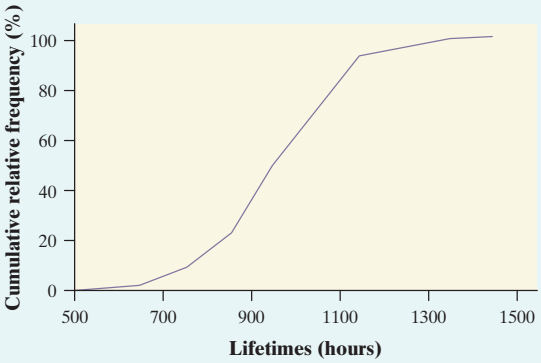


(a) Estimate the interquartile range of this distribution. Show your method.

(b) What is the percentile for the shopper who spent $19.50?

(c) Draw the histogram that corresponds to this graph.

10. **Light it up!** The graph below is a cumulative relative frequency graph showing the lifetimes (in hours) of 200 lamps.[4]



(a) Estimate the 60th percentile of this distribution. Show your method.

(b) What is the percentile for a lamp that lasted 900 hours?

(c) Draw a histogram that corresponds to this graph.

11. SAT versus ACT Eleanor scores 680 on the SAT Mathematics test. The distribution of SAT scores is symmetric and single-peaked, with mean 500 and standard deviation 100. Gerald takes the American College Testing (ACT) Mathematics test and scores 27. ACT scores also follow a symmetric, single-peaked distribution—but with mean 18 and standard deviation 6. Find the standardized scores for both students. Assuming that both tests measure the same kind of ability, who has the higher score?

pg 91

12. Comparing batting averages Three landmarks of baseball achievement are Ty Cobb's batting average of 0.420 in 1911, Ted Williams's 0.406 in 1941, and George Brett's 0.390 in 1980. These batting averages cannot be compared directly because the distribution of major league batting averages has changed over the years. The distributions are quite symmetric, except for outliers such as Cobb, Williams, and Brett. While the mean batting average has been held roughly constant by rule changes and the balance between hitting and pitching, the standard deviation has dropped over time. Here are the facts:

Decade	Mean	Standard deviation
1910s	0.266	0.0371
1940s	0.267	0.0326
1970s	0.261	0.0317

Find the standardized scores for Cobb, Williams, and Brett. Who was the best hitter?[5]

13. Measuring bone density Individuals with low bone density have a high risk of broken bones (fractures). Physicians who are concerned about low bone density (osteoporosis) in patients can refer them for specialized testing. Currently, the most common method for testing bone density is dual-energy X-ray absorptiometry (DEXA). A patient who undergoes a DEXA test usually gets bone density results in grams per square centimeter (g/cm^2) and in standardized units.

Judy, who is 25 years old, has her bone density measured using DEXA. Her results indicate a bone density in the hip of 948 g/cm^2 and a standardized score of $z = -1.45$. In the reference population of 25-year-old women like Judy, the mean bone density in the hip is 956 g/cm^2.[6]

(a) Judy has not taken a statistics class in a few years. Explain to her in simple language what the standardized score tells her about her bone density.

(b) Use the information provided to calculate the standard deviation of bone density in the reference population.

14. Comparing bone density Refer to the previous exercise. One of Judy's friends, Mary, has the bone density in her hip measured using DEXA. Mary is 35 years old. Her bone density is also reported as 948 g/cm^2, but her standardized score is $z = 0.50$. The mean bone density in the hip for the reference population of 35-year-old women is 944 grams/cm^2.

(a) Whose bones are healthier—Judy's or Mary's? Justify your answer.

(b) Calculate the standard deviation of the bone density in Mary's reference population. How does this compare with your answer to Exercise 13(b)? Are you surprised?

Exercises 15 and 16 refer to the dotplot and summary statistics of salaries for players on the World Champion 2008 Philadelphia Phillies baseball team.[7]

Variable	n	Mean	Std. dev.	Min	Q_1	Med	Q_3	Max
Salary	29	3388617	3767484	390000	440000	1400000	6000000	14250000

15. Baseball salaries Brad Lidge played a crucial role as the Phillies' "closer," pitching the end of many games throughout the season. Lidge's salary for the 2008 season was $6,350,000.

pg 90

(a) Find the percentile corresponding to Lidge's salary. Explain what this value means.

(b) Find the z-score corresponding to Lidge's salary. Explain what this value means.

16. Baseball salaries Did Ryan Madson, who was paid $1,400,000, have a high salary or a low salary compared with the rest of the team? Justify your answer by calculating and interpreting Madson's percentile and z-score.

17. The scores on Ms. Martin's statistics quiz had a mean of 12 and a standard deviation of 3. Ms. Martin wants to transform the scores to have a mean of 75 and a standard deviation of 12. What transformations should she apply to each test score? Explain.

18. Mr. Olsen uses an unusual grading system in his class. After each test, he transforms the scores to have a mean of 0 and a standard deviation of 1. Mr. Olsen then assigns a grade to each student based on the transformed score. On his most recent test, the class's scores had a mean of 68 and a standard deviation of 15. What transformations should he apply to each test score? Explain.

19. **Tall or short?** Mr. Walker measures the heights (in pg 95 inches) of the students in one of his classes. He uses a computer to calculate the following numerical summaries:

Mean	Std. dev.	Min	Q_1	Med	Q_3	Max
69.188	3.20	61.5	67.75	69.5	71	74.5

Next, Mr. Walker has his entire class stand on their chairs, which are 18 inches off the ground. Then he measures the distance from the top of each student's head to the floor.

(a) Find the mean and median of these measurements. Show your work.

(b) Find the standard deviation and *IQR* of these measurements. Show your work.

20. **Teacher raises** A school system employs teachers at salaries between $28,000 and $60,000. The teachers' union and the school board are negotiating the form of next year's increase in the salary schedule.

(a) If every teacher is given a flat $1000 raise, what will this do to the mean salary? To the median salary? Explain your answers.

(b) What would a flat $1000 raise do to the extremes and quartiles of the salary distribution? To the standard deviation of teachers' salaries? Explain your answers.

21. **Tall or short?** Refer to Exercise 19. Mr. Walker converts his students' original heights from inches to feet.

(a) Find the mean and median of the students' heights in feet. Show your work.

(b) Find the standard deviation and *IQR* of the students' heights in feet. Show your work.

22. **Teacher raises** Refer to Exercise 20. If each teacher receives a 5% raise instead of a flat $1000 raise, the amount of the raise will vary from $1400 to $3000, depending on the present salary.

(a) What will this do to the mean salary? To the median salary? Explain your answers.

(b) Will a 5% raise increase the *IQR*? Will it increase the standard deviation? Explain your answers.

23. **Cool pool?** Coach Ferguson uses a thermometer to measure the temperature (in degrees Celsius) at 20 different locations in the school swimming pool. An analysis of the data yields a mean of 25°C and a standard deviation of 2°C. Find the mean and standard deviation of the temperature readings in degrees Fahrenheit (recall that °F = (9/5)°C + 32).

24. **Measure up** Clarence measures the diameter of each tennis ball in a bag with a standard ruler. Unfortunately, he uses the ruler incorrectly so that each of his measurements is 0.2 inches too large. Clarence's data had a mean of 3.2 inches and a standard deviation of 0.1 inches. Find the mean and standard deviation of the corrected measurements in centimeters (recall that 1 inch = 2.54 cm).

Multiple choice: Select the best answer for Exercises 25 to 30.

25. Jorge's score on Exam 1 in his statistics class was at the 64th percentile of the scores for all students. His score falls

(a) between the minimum and the first quartile.

(b) between the first quartile and the median.

(c) between the median and the third quartile.

(d) between the third quartile and the maximum.

(e) at the mean score for all students.

26. When Sam goes to a restaurant, he always tips the server $2 plus 10% of the cost of the meal. If Sam's distribution of meal costs has a mean of $9 and a standard deviation of $3, what are the mean and standard deviation of the distribution of his tips?

(a) $2.90, $0.30

(b) $2.90, $2.30

(c) $9.00, $3.00

(d) $11.00, $2.00

(e) $2.00, $0.90

27. Scores on the ACT college entrance exam follow a bell-shaped distribution with mean 18 and standard deviation 6. Wayne's standardized score on the ACT was −0.5. What was Wayne's actual ACT score?

(a) 5.5 (b) 12 (c) 15 (d) 17.5 (e) 21

28. George has an average bowling score of 180 and bowls in a league where the average for all bowlers is 150 and the standard deviation is 20. Bill has an average bowling score of 190 and bowls in a league where the average is 160 and the standard deviation is 15. Who ranks higher in his own league, George or Bill?

(a) Bill, because his 190 is higher than George's 180.

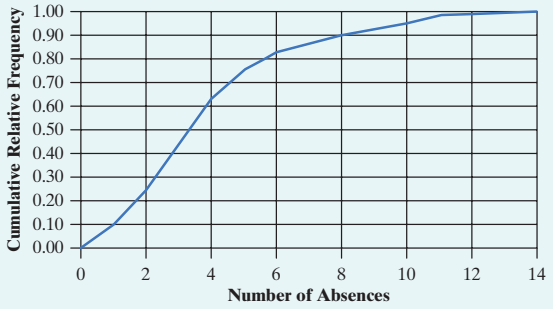
(b) Bill, because his standardized score is higher than George's.

(c) Bill and George have the same rank in their leagues, because both are 30 pins above the mean.

(d) George, because his standardized score is higher than Bill's.

(e) George, because the standard deviation of bowling scores is higher in his league.

Exercises 29 and 30 refer to the following setting. The number of absences during the fall semester was recorded for each student in a large elementary school. The distribution of absences is displayed in the following cumulative relative frequency graph.



29. What is the interquartile range (*IQR*) for the distribution of absences?

(a) 1 (c) 3 (e) 14

(b) 2 (d) 5

30. If the distribution of absences was displayed in a histogram, what would be the best description of the histogram's shape?

(a) Symmetric

(b) Uniform

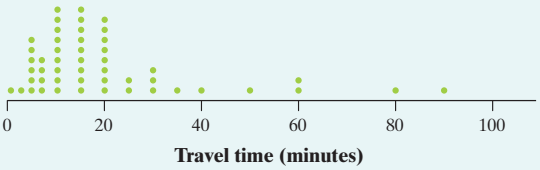
(c) Skewed left

(d) Skewed right

(e) Cannot be determined

Exercises 31 and 32 refer to the following setting. We used CensusAtSchool's Random Data Selector to choose a sample of 50 Canadian students who completed a survey in a recent year.

31. Travel time (1.2) The dotplot below displays data on students' responses to the question "How long does it usually take you to travel to school?" Describe the shape, center, and spread of the distribution. Are there any outliers?



32. Lefties (1.1) Students were asked, "Are you right-handed, left-handed, or ambidextrous?" The responses are shown below (R = right-handed; L = left-handed; A = ambidextrous).

R	R	R	R	R	R	R	R	R	R	R	L	R	R
R	R	R	R	R	R	R	R	R	R	R	R	R	A
R	R	R	R	A	R	R	L	R	R	R	R	L	A
R	R	R	R	R	R	R	R						

(a) Make an appropriate graph to display these data.

(b) Over 10,000 Canadian high school students took the CensusAtSchool survey that year. What percent of this population would you estimate is left-handed? Justify your answer.

2.2 Density Curves and Normal Distributions

WHAT YOU WILL LEARN

By the end of the section, you should be able to:

- Estimate the relative locations of the median and mean on a density curve.
- Use the 68–95–99.7 rule to estimate areas (proportions of values) in a Normal distribution.
- Use Table A or technology to find (i) the proportion of *z*-values in a specified interval, or (ii) a *z*-score from a percentile in the standard Normal distribution.
- Use Table A or technology to find (i) the proportion of values in a specified interval, or (ii) the value that corresponds to a given percentile in any Normal distribution.
- Determine whether a distribution of data is approximately Normal from graphical and numerical evidence.

In Chapter 1, we developed a kit of graphical and numerical tools for describing distributions. Our work gave us a clear strategy for exploring data from a single quantitative variable.

EXPLORING QUANTITATIVE DATA

1. Always plot your data: make a graph, usually a dotplot, stemplot, or histogram.
2. Look for the overall pattern (shape, center, spread) and for striking departures such as outliers.
3. Calculate numerical summaries to briefly describe center and spread.

In this section, we add one more step to this strategy.

4. Sometimes the overall pattern of a large number of observations is so regular that we can describe it by a smooth curve.

Density Curves

Figure 2.7 is a histogram of the scores of all 947 seventh-grade students in Gary, Indiana, on the vocabulary part of the Iowa Test of Basic Skills (ITBS).[8] Scores on this national test have a very regular distribution. The histogram is symmetric, and both tails fall off smoothly from a single center peak. There are no large gaps or obvious outliers. The smooth curve drawn through the tops of the histogram bars in Figure 2.7 is a good description of the overall pattern of the data.

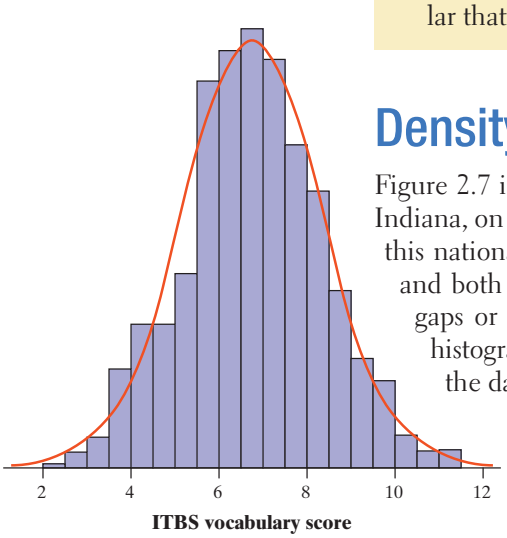


FIGURE 2.7 Histogram of the Iowa Test of Basic Skills (ITBS) vocabulary scores of all seventh-grade students in Gary, Indiana. The smooth curve shows the overall shape of the distribution.

EXAMPLE

Seventh-Grade Vocabulary Scores

From histogram to density curve

Our eyes respond to the areas of the bars in a histogram. The bar areas represent relative frequencies (proportions) of the observations. Figure 2.8(a) is a copy of Figure 2.7 with the leftmost bars shaded. The area of the shaded bars in Figure 2.8(a) represents the proportion of students with vocabulary scores less than 6.0. There are 287 such students, who make up the proportion 287/947 = 0.303 of all Gary seventh-graders. In other words, a score of 6.0 corresponds to about the 30th percentile.

The total area of the bars in the histogram is 100% (a proportion of 1), because all of the observations are represented. Now look at the curve drawn through the tops of the bars. In Figure 2.8(b), the area under the curve to the left of 6.0 is shaded. In moving from histogram bars to a smooth curve, we make a specific choice: adjust the scale of the graph so that *the total area under the curve is exactly 1*. Now

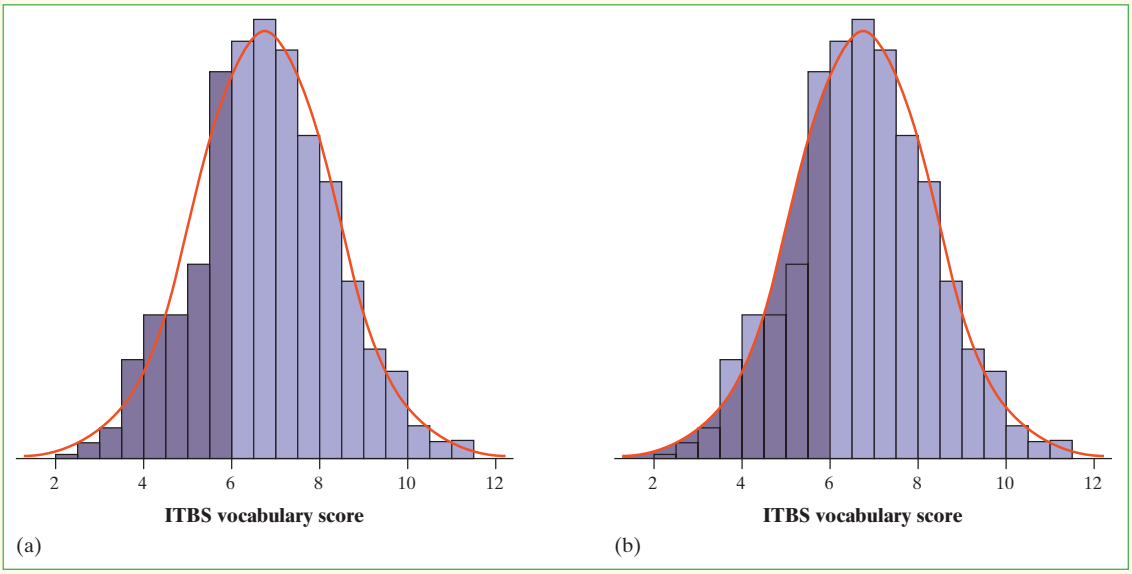


FIGURE 2.8 (a) The proportion of scores less than or equal to 6.0 in the actual data is 0.303. (b) The proportion of scores less than or equal to 6.0 from the density curve is 0.293.

the total area represents all the observations, just like with the histogram. We can then interpret areas under the curve as proportions of the observations.

The shaded area under the curve in Figure 2.8(b) represents the proportion of students with scores lower than 6.0. This area is 0.293, only 0.010 away from the actual proportion 0.303. So our estimate based on the curve is that a score of 6.0 falls at about the 29th percentile. You can see that areas under the curve give good approximations to the actual distribution of the 947 test scores. In practice, it might be easier to use this curve to estimate relative frequencies than to determine the actual proportion of students by counting data values.

A curve like the one in the previous example is called a **density curve.**

DEFINITION: Density curve

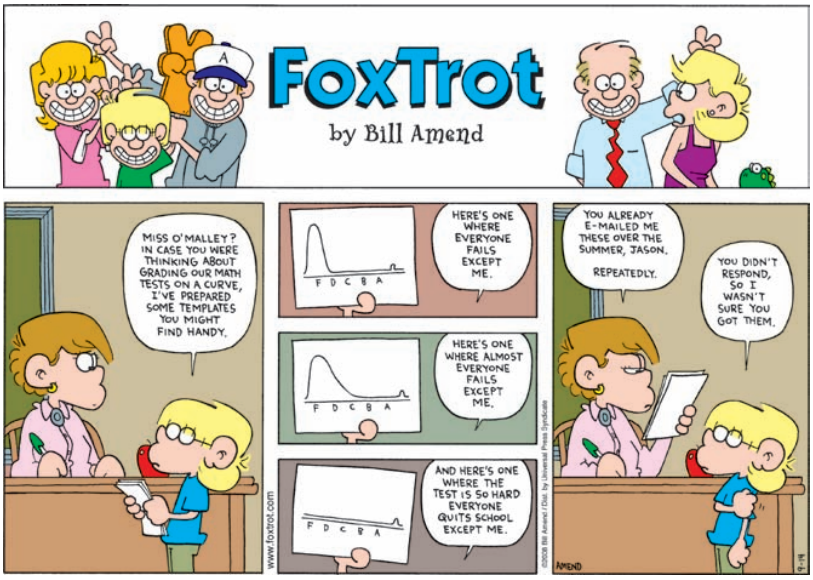
A **density curve** is a curve that

- is always on or above the horizontal axis, and
- has area exactly 1 underneath it.

A density curve describes the overall pattern of a distribution. The area under the curve and above any interval of values on the horizontal axis is the proportion of all observations that fall in that interval.

Density curves, like distributions, come in many shapes. A density curve is often a good description of the overall pattern of a distribution. Outliers, which are departures from the overall pattern, are not described by the curve.

No set of real data is exactly described by a density curve. The curve is an approximation that is easy to use and accurate enough for practical use.



Describing Density Curves

Our measures of center and spread apply to density curves as well as to actual sets of observations. Areas under a density curve represent proportions of the total number of observations. The median of a data set is the point with half the observations on either side. So the **median of a density curve** is the "equal-areas point," the point with half the area under the curve to its left and the remaining half of the area to its right.

Because density curves are idealized patterns, a symmetric density curve is exactly symmetric. The median of a symmetric density curve is therefore at its center. Figure 2.9(a) shows a symmetric density curve with the median marked. It isn't so easy to spot the equal-areas point on a skewed curve. There are mathematical ways of finding the median for any density curve. That's how we marked the median on the skewed curve in Figure 2.9(b).

What about the mean? The mean of a set of observations is their arithmetic average. As we saw in Chapter 1, the mean is also the "balance point" of a distribution. That is, if we think of the observations as weights strung out along a thin rod, the mean is the point at which the rod would balance. This fact is also true of

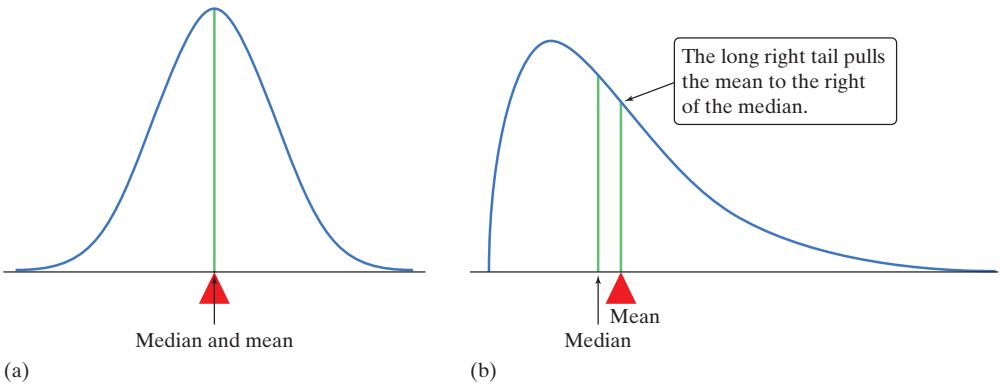


FIGURE 2.9 (a) The median and mean of a symmetric density curve both lie at the center of symmetry. (b) The median and mean of a right-skewed density curve. The mean is pulled away from the median toward the long tail.

density curves. The **mean of a density curve** is the point at which the curve would balance if made of solid material. Figure 2.10 illustrates this fact about the mean.

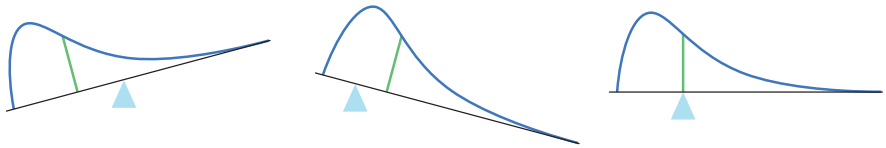


FIGURE 2.10 The mean is the balance point of a density curve.

A symmetric curve balances at its center because the two sides are identical. *The mean and median of a symmetric density curve are equal,* as in Figure 2.9(a). We know that the mean of a skewed distribution is pulled toward the long tail. Figure 2.9(b) shows how the mean of a skewed density curve is pulled toward the long tail more than the median is.

DISTINGUISHING THE MEDIAN AND MEAN OF A DENSITY CURVE

The **median** of a density curve is the equal-areas point, the point that divides the area under the curve in half.

The **mean** of a density curve is the balance point, at which the curve would balance if made of solid material.

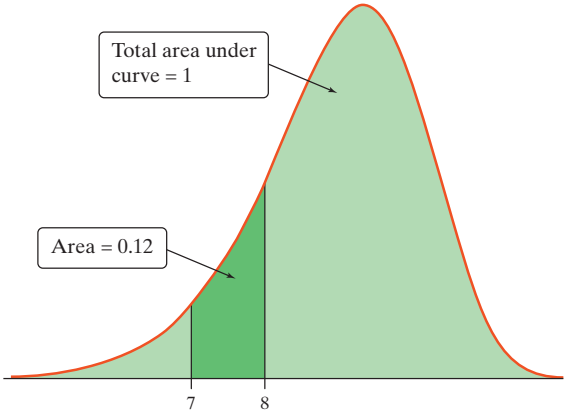
The median and mean are the same for a symmetric density curve. They both lie at the center of the curve. The mean of a skewed curve is pulled away from the median in the direction of the long tail.

You probably noticed that we used the same notation for the mean and standard deviation of a population in Chapter 1, μ and σ, as we do here for the mean and standard deviation of a density curve.

Because a density curve is an idealized description of a distribution of data, we distinguish between the mean and standard deviation of the density curve and the mean $\bar{x}$ and standard deviation s_x computed from the actual observations. The usual notation for the mean of a density curve is μ (the Greek letter mu). We write the standard deviation of a density curve as σ (the Greek letter sigma). We can roughly locate the mean μ of any density curve by eye, as the balance point. There is no easy way to locate the standard deviation σ by eye for density curves in general.

CHECK YOUR UNDERSTANDING

Use the figure shown to answer the following questions.



1. Explain why this is a legitimate density curve.

2. About what proportion of observations lie between 7 and 8?

3. Trace the density curve onto your paper. Mark the approximate location of the median.

4. Now mark the approximate location of the mean. Explain why the mean and median have the relationship that they do in this case.

Normal Distributions

One particularly important class of density curves has already appeared in Figures 2.7, 2.8, and 2.9(a). They are called **Normal curves.** The distributions they describe are called **Normal distributions.** Normal distributions play a large role in statistics, but they are rather special and not at all "normal" in the sense of being usual or typical. We capitalize Normal to remind you that these curves are special.

Look at the two Normal curves in Figure 2.11. They illustrate several important facts:

- All Normal curves have the same overall shape: symmetric, single-peaked, and bell-shaped.
- Any specific Normal curve is completely described by giving its mean μ and its standard deviation σ.

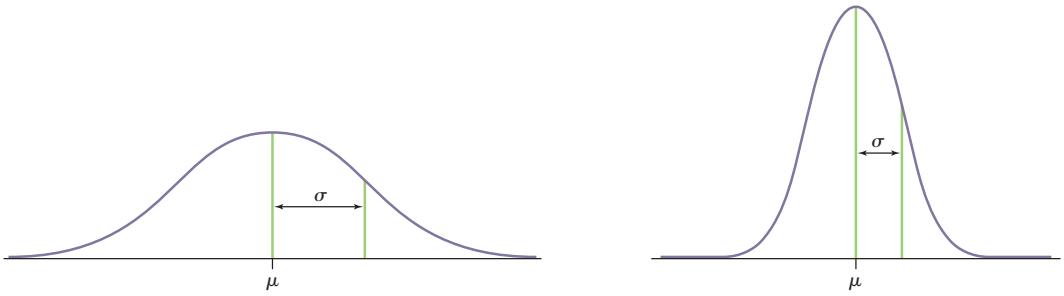


FIGURE 2.11 Two Normal curves, showing the mean μ and standard deviation σ.

- The mean is located at the center of the symmetric curve and is the same as the median. Changing μ without changing σ moves the Normal curve along the horizontal axis without changing its spread.
- The standard deviation σ controls the spread of a Normal curve. Curves with larger standard deviations are more spread out.

The standard deviation σ is the natural measure of spread for Normal distributions. Not only do μ and σ completely determine the shape of a Normal curve, but we can locate σ by eye on a Normal curve. Here's how. Imagine that you are skiing down a mountain that has the shape of a Normal curve. At first, you descend at an ever-steeper angle as you go out from the peak:

Fortunately, before you find yourself going straight down, the slope begins to grow flatter rather than steeper as you go out and down:

The points at which this change of curvature takes place are located at a distance σ on either side of the mean μ. (Advanced math students know these points as "inflection points.") You can feel the change as you run a pencil along a Normal curve and so find the standard deviation. *Remember that μ and σ alone do not specify the shape of most distributions.* The shape of density curves in general does not reveal σ. These are special properties of Normal distributions.

DEFINITION: Normal distribution and Normal curve

A **Normal distribution** is described by a Normal density curve. Any particular Normal distribution is completely specified by two numbers: its mean μ and standard deviation σ. The mean of a Normal distribution is at the center of the symmetric **Normal curve**. The standard deviation is the distance from the center to the change-of-curvature points on either side.

We abbreviate the Normal distribution with mean μ and standard deviation σ as $N(\mu, \sigma)$.

Normal curves were first applied to data by the great mathematician Carl Friedrich Gauss (1777–1855). He used them to describe the small errors made by astronomers and surveyors in repeated careful measurements of the same quantity. You will sometimes see Normal distributions labeled "Gaussian" in honor of Gauss.

Why are the Normal distributions important in statistics? Here are three reasons. First, Normal distributions are good descriptions for some distributions of *real data*. Distributions that are often close to Normal include

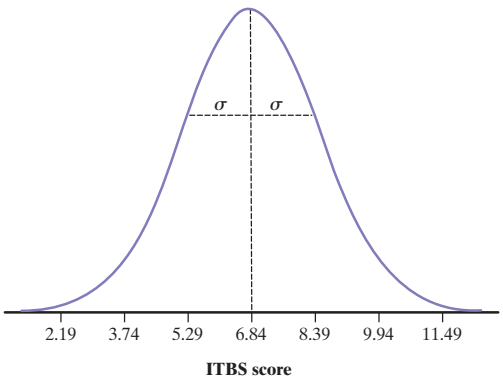
- scores on tests taken by many people (such as SAT exams and IQ tests),
- repeated careful measurements of the same quantity (like the diameter of a tennis ball), and
- characteristics of biological populations (such as lengths of crickets and yields of corn).

Second, Normal distributions are good approximations to the results of many kinds of *chance outcomes*, like the number of heads in many tosses of a fair coin. Third, and most important, we will see that many *statistical inference* procedures are based on Normal distributions.

Even though many sets of data follow a Normal distribution, many do not. Most income distributions, for example, are skewed to the right and so are not Normal. Some distributions are symmetric but not Normal or even close to Normal. The uniform distribution of Exercise 35 (page 128) is one such example. Non-Normal data, like non-normal people, not only are common but are sometimes more interesting than their Normal counterparts.

The 68–95–99.7 Rule

Earlier, we saw that the distribution of Iowa Test of Basic Skills (ITBS) vocabulary scores for seventh-grade students in Gary, Indiana, is symmetric, single-peaked, and bell-shaped. Suppose that the distribution of scores over time is exactly Normal with mean $\mu = 6.84$ and standard deviation $\sigma = 1.55$. (These are the mean and standard deviation of the 947 actual scores.) The figure shows the Normal density curve for this distribution with the points 1, 2, and 3 standard deviations from the mean labeled on the horizontal axis.



How unusual is it for a Gary seventh-grader to get an ITBS score above 9.94? As the following activity shows, the answer to this question is surprisingly simple.

ACTIVITY | The Normal density curve applet

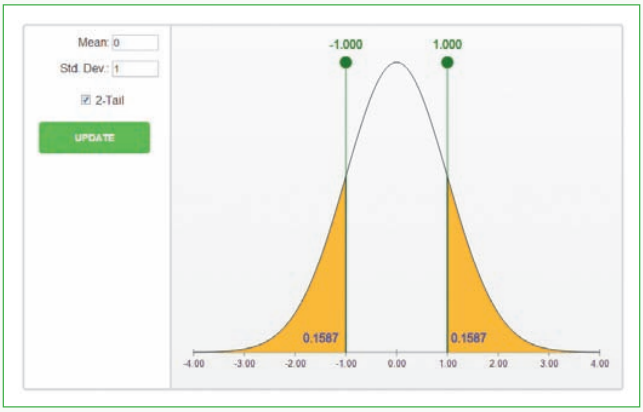
MATERIALS:
Computer with Internet access

In this Activity, you will use the *Normal Density Curve* applet at the book's Web site (www.whfreeman.com/tps5e) to explore an interesting property of Normal distributions. A graph similar to what you will see when you launch the applet is shown below. The applet finds the area under the curve in the region indicated by the green flags. If you drag one flag past the other, the applet will show the area under the curve between the two flags. When the "2-Tail" box is checked, the applet calculates symmetric areas around the mean.

Use the applet to help you answer the following questions.

1. If you put one flag at the extreme left of the curve and the second flag exactly in the middle, what proportion is reported by the applet? Why does this value make sense?

2. If you place the two flags exactly one standard deviation on either side of the mean, what does the applet say is the area between them?



3. What percent of the area under the Normal curve lies within 2 standard deviations of the mean?

4. Use the applet to show that about 99.7% of the area under the Normal density curve lies within three standard deviations of the mean. Does this mean that about 99.7%/2 = 49.85% will lie within one and a half standard deviations? Explain.

5. Change the mean to 100 and the standard deviation to 15. Then click "Update." What percent of the area under this Normal density curve lies within one, two, and three standard deviations of the mean?

6. Change the mean to 6.84 and the standard deviation to 1.55. (These values are from the ITBS vocabulary scores in Gary, Indiana.) Answer the question from Step 5.

7. *Summarize:* Complete the following sentence: "For any Normal density curve, the area under the curve within one, two, and three standard deviations of the mean is about ___%, ___%, and ___%."

Although there are many Normal curves, they all have properties in common. In particular, all Normal distributions obey the following rule.

DEFINITION: The 68–95–99.7 rule

In a Normal distribution with mean μ and standard deviation σ:

- Approximately **68%** of the observations fall within σ of the mean μ.
- Approximately **95%** of the observations fall within 2σ of the mean μ.
- Approximately **99.7%** of the observations fall within 3σ of the mean μ.

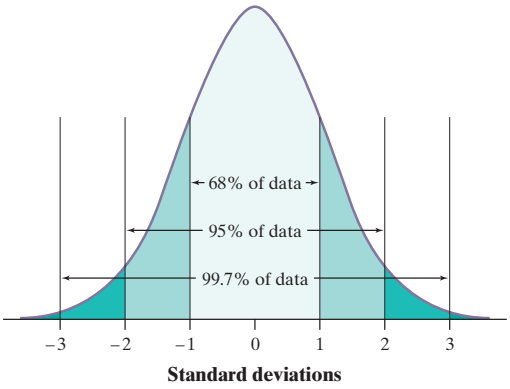


Figure 2.12 illustrates the 68–95–99.7 rule. (Some people refer to this result as the "empirical rule.") By remembering these three numbers, you can think about Normal distributions without constantly making detailed calculations.

Here's an example that shows how we can use the 68–95–99.7 rule to estimate the percent of observations in a specified interval.

FIGURE 2.12 The 68–95–99.7 rule for Normal distributions.

EXAMPLE ITBS Vocabulary Scores

Using the 68–95–99.7 rule

PROBLEM: The distribution of ITBS vocabulary scores for seventh-graders in Gary, Indiana, is $N(6.84, 1.55)$.

(a) What percent of the ITBS vocabulary scores are less than 3.74? Show your work.

(b) What percent of the scores are between 5.29 and 9.94? Show your work.

SOLUTION:

(a) Notice that a score of 3.74 is exactly two standard deviations below the mean. By the 68–95–99.7 rule, about 95% of all scores are between

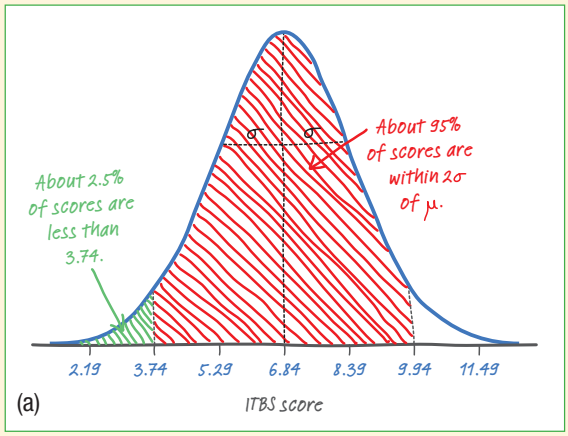
$$\mu - 2\sigma = 6.84 - (2)(1.55) = 6.84 - 3.10 = 3.74$$

and

$$\mu + 2\sigma = 6.84 + (2)(1.55) = 6.84 + 3.10 = 9.94$$

The other 5% of scores are outside this range. Because Normal distributions are symmetric, half of these scores are lower than 3.74 and half are higher than 9.94. That is, about 2.5% of the ITBS scores are below 3.74. Figure 2.13(a) shows this reasoning in picture form.

(b) Let's start with a picture. Figure 2.13(b) shows the area under the Normal density curve between 5.29 and 9.94. We can see that about 68% + 13.5% = 81.5% of ITBS scores are between 5.29 and 9.94.



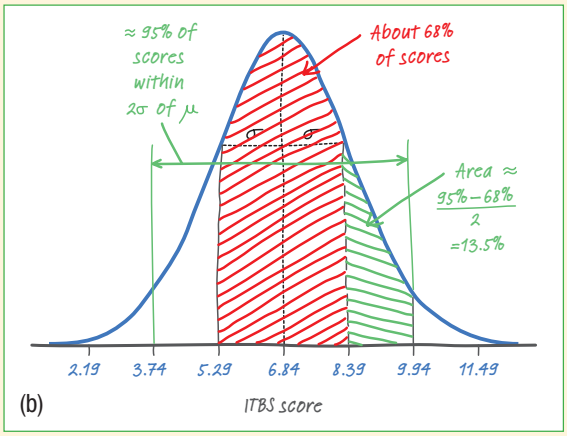


FIGURE 2.13 (a) Finding the percent of Iowa Test scores less than 3.74. (b) Finding the percent of Iowa Test scores between 5.29 and 9.94.

For Practice *Try Exercise* 43

Chebyshev's inequality is an interesting result, but it is not required for the AP® Statistics exam.

The 68–95–99.7 rule applies *only* to Normal distributions. Is there a similar rule that would apply to *any* distribution? Sort of. A result known as **Chebyshev's inequality** says that in any distribution, the proportion of observations falling within k standard deviations of the mean is *at least* $1 - \frac{1}{k^2}$. If $k = 2$, for example, Chebyshev's inequality tells us that at least $1 - \frac{1}{2^2} = 0.75$ of the observations in *any* distribution are within 2 standard deviations of the mean. For Normal distributions, we know that this proportion is much higher than 0.75. In fact, it's approximately 0.95.

THINK ABOUT IT

All models are wrong, but some are useful! The 68–95–99.7 rule describes distributions that are exactly Normal. Real data such as the actual ITBS scores are never exactly Normal. For one thing, ITBS scores are reported only to the nearest tenth. A score can be 9.9 or 10.0 but not 9.94. We use a Normal distribution because it's a good approximation, and because we think of the knowledge that the test measures as continuous rather than stopping at tenths.

How well does the 68–95–99.7 rule describe the actual ITBS scores? Well, 900 of the 947 scores are between 3.74 and 9.94. That's 95.04%, very accurate indeed. Of the remaining 47 scores, 20 are below 3.74 and 27 are above 9.94. The tails of the actual data are not quite equal, as they would be in an exactly Normal distribution. Normal distributions often describe real data better in the center of the distribution than in the extreme high and low tails.

As famous statistician George Box once noted, "All models are wrong, but some are useful!"

CHECK YOUR UNDERSTANDING

The distribution of heights of young women aged 18 to 24 is approximately N(64.5, 2.5).

1. Sketch a Normal density curve for the distribution of young women's heights. Label the points one, two, and three standard deviations from the mean.
2. What percent of young women have heights greater than 67 inches? Show your work.
3. What percent of young women have heights between 62 and 72 inches? Show your work.

The Standard Normal Distribution

As the 68–95–99.7 rule suggests, all Normal distributions share many properties. In fact, all Normal distributions are the same if we measure in units of size σ from the mean μ as center. Changing to these units requires us to standardize, just as we did in Section 2.1:

$$z = \frac{x - \mu}{\sigma}$$

If the variable we standardize has a Normal distribution, then so does the new variable z. (Recall that subtracting a constant and dividing by a constant don't

change the shape of a distribution.) This new distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ is called the **standard Normal distribution**.

DEFINITION: Standard Normal distribution

The **standard Normal distribution** is the Normal distribution with mean 0 and standard deviation 1 (Figure 2.14).

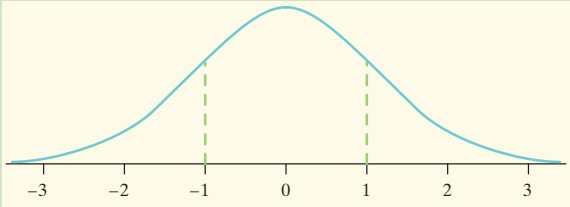


FIGURE 2.14 The standard Normal distribution.

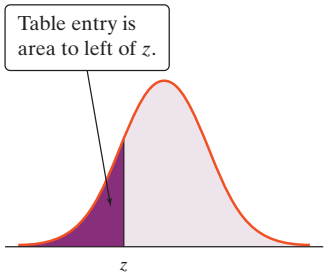
If a variable x has any Normal distribution $N(\mu, \sigma)$ with mean μ and standard deviation σ, then the standardized variable

$$z = \frac{x - \mu}{\sigma}$$

has the standard Normal distribution $N(0,1)$.

An area under a density curve is a proportion of the observations in a distribution. Any question about what proportion of observations lies in some range of values can be answered by finding an area under the curve. In a standard Normal distribution, the 68–95–99.7 rule tells us that about 68% of the observations fall between $z = -1$ and $z = 1$ (that is, within one standard deviation of the mean). What if we want to find the percent of observations that fall between $z = -1.25$ and $z = 1.25$? The 68–95–99.7 rule can't help us.

Because all Normal distributions are the same when we standardize, we can find areas under any Normal curve from a single table, a table that gives areas under the curve for the standard Normal distribution. Table A, the **standard Normal table,** gives areas under the standard Normal curve. You can find Table A in the back of the book.



DEFINITION: The standard Normal table

Table A is a table of areas under the standard Normal curve. The table entry for each value z is the area under the curve to the left of z.

z	.00	.01	.02
0.7	.7580	.7611	.7642
0.8	.7881	.7910	.7939
0.9	.8159	.8186	.8212

For instance, suppose we wanted to find the proportion of observations from the standard Normal distribution that are less than 0.81. To find the area to the left of $z = 0.81$, locate 0.8 in the left-hand column of Table A, then locate the remaining digit 1 as .01 in the top row. The entry opposite 0.8 and under .01 is .7910. This is the area we seek. A reproduction of the relevant portion of Table A is shown in the margin.

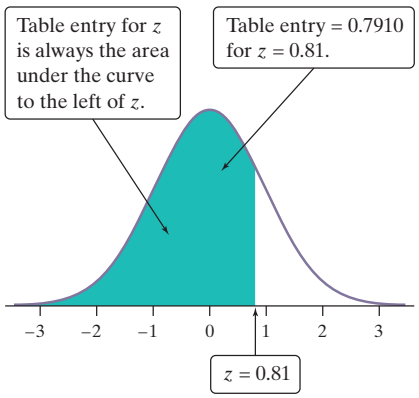


FIGURE 2.15 The area under a standard Normal curve to the left of the point $z = 0.81$ is 0.7910.

Figure 2.15 illustrates the relationship between the value $z = 0.81$ and the area 0.7910. *Note that we have made a connection between z-scores and percentiles when the shape of a distribution is Normal.*

EXAMPLE Standard Normal Distribution

Finding area to the right

What if we wanted to find the proportion of observations from the standard Normal distribution that are *greater than* −1.78? To find the area to the right of $z = -1.78$, locate −1.7 in the left-hand column of Table A, then locate the remaining digit 8 as .08 in the top row. The corresponding entry is .0375. (See the excerpt from Table A in the margin.)

z	.07	.08	.09
−1.8	.0307	.0301	.0294
−1.7	.0384	.0375	.0367
−1.6	.0475	.0465	.0455

This is the area *to the left* of $z = -1.78$. To find the area *to the right* of $z = -1.78$, we use the fact that the total area under the standard Normal density curve is 1. So the desired proportion is $1 - 0.0375 = 0.9625$.

Figure 2.16 illustrates the relationship between the value $z = -1.78$ and the area 0.9625.

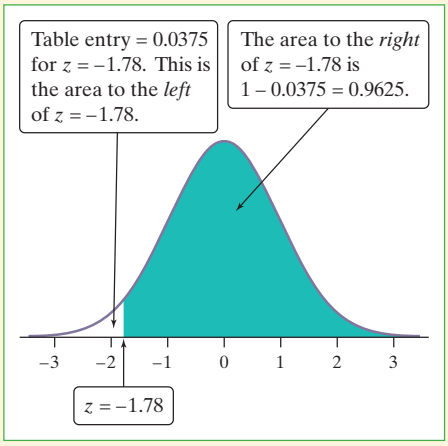


FIGURE 2.16 The area under a standard Normal curve to the right of the point $z = -1.78$ is 0.9625.

A common student mistake is to look up a z-value in Table A and report the entry corresponding to that z-value, regardless of whether the problem asks for the area to the left or to the right of that z-value. To prevent making this mistake, always sketch the standard Normal curve, mark the z-value, and shade the area of interest. And before you finish, make sure your answer is reasonable in the context of the problem.

EXAMPLE Catching Some "z"s

Finding areas under the standard Normal curve

PROBLEM: Find the proportion of observations from the standard Normal distribution that are between -1.25 and 0.81.

SOLUTION: From Table A, the area to the left of $z = 0.81$ is 0.7910 and the area to the left of $z = -1.25$ is 0.1056. So the area under the standard Normal curve between these two z-scores is $0.7910 - 0.1056 = 0.6854$. Figure 2.17 shows why this approach works.

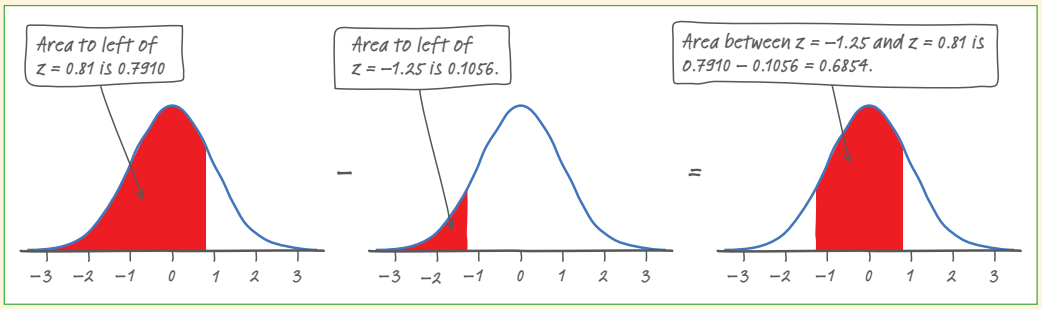
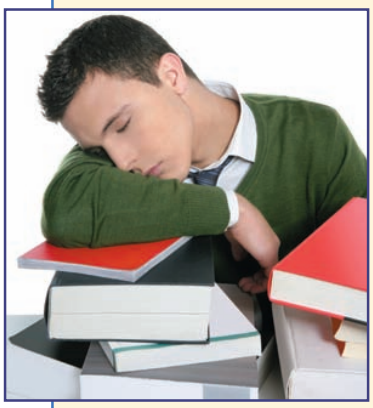


FIGURE 2.17 One way to find the area between $z = -1.25$ and $z = 0.81$ under the standard Normal curve.



Here's another way to find the desired area. The area to the left of $z = -1.25$ under the standard Normal curve is 0.1056. The area to the *right* of $z = 0.81$ is $1 - 0.7910 = 0.2090$. So the area *between* these two z-scores is

$$1 - (0.1056 + 0.2090) = 1 - 0.3146 = 0.6854$$

Figure 2.18 shows this approach in picture form.

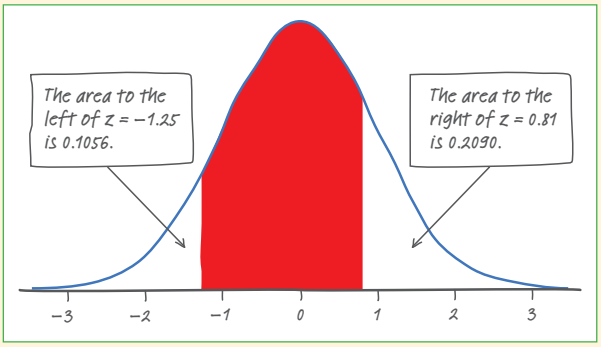


FIGURE 2.18 The area under the standard Normal curve between $z = -1.25$ and $z = 0.81$ is 0.6854.

For Practice *Try Exercise* 49

Working backward: From areas to *z*-scores: So far, we have used Table A to find areas under the standard Normal curve from *z*-scores. What if we want to find the *z*-score that corresponds to a particular area? For example, let's find the 90th percentile of the standard Normal curve. We're looking for the *z*-score that has 90% of the area to its left, as shown in Figure 2.19.

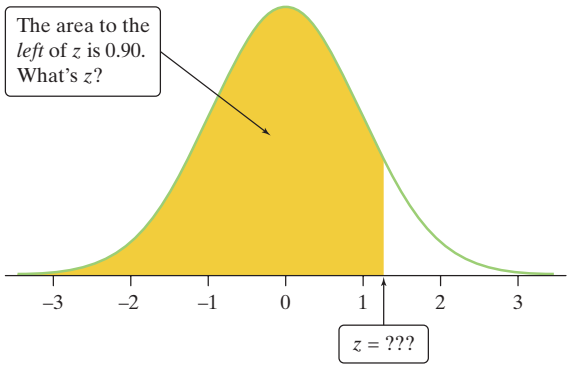


FIGURE 2.19 The *z*-score with area 0.90 to its left under the standard Normal curve.

z	.07	.08	.09
1.1	.8790	.8810	.8830
1.2	.8980	.8997	.9015
1.3	.9147	.9162	.9177

Because Table A gives areas to the left of a specified *z*-score, all we need to do is find the value closest to 0.90 in the middle of the table. From the reproduced portion of Table A, you can see that the desired *z*-score is $z = 1.28$. That is, the area to the left of $z = 1.28$ is approximately 0.90.

CHECK YOUR UNDERSTANDING

Use Table A in the back of the book to find the proportion of observations from a standard Normal distribution that fall in each of the following regions. In each case, sketch a standard Normal curve and shade the area representing the region.

1. $z < 1.39$
2. $z > -2.15$
3. $-0.56 < z < 1.81$

Use Table A to find the value *z* from the standard Normal distribution that satisfies each of the following conditions. In each case, sketch a standard Normal curve with your value of *z* marked on the axis.

4. The 20th percentile
5. 45% of all observations are greater than *z*

You can use the *Normal Density Curve* applet at www.whfreeman.com/tps5e to confirm areas under the standard Normal curve. Just enter mean 0 and standard deviation 1, and then drag the flags to the appropriate locations. Of course, you can also confirm Normal curve areas with your calculator.

5. TECHNOLOGY CORNER FROM *z*-SCORES TO AREAS, AND VICE VERSA

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

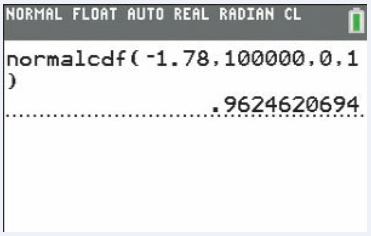
Finding areas: The `normalcdf` command on the TI-83/84 (`normCdf` on the TI-89) can be used to find areas under a Normal curve. The syntax is `normalcdf(lower bound,upper bound,mean,standard deviation)`. Let's use this command to confirm our answers to the previous two examples.

1. What proportion of observations from the standard Normal distribution are greater than −1.78?

Recall that the standard Normal distribution has mean 0 and standard deviation 1.

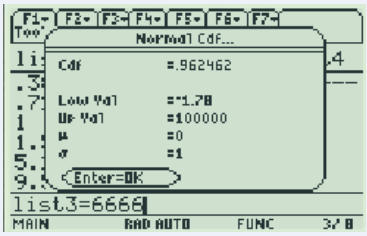
TI-83/84

- Press 2nd VARS (Distr) and choose normalcdf(. **OS 2.55 or later:** In the dialog box, enter these values: lower:−1.78, upper:100000, μ:0, σ:1, choose Paste, and then press ENTER. **Older OS:** Complete the command normalcdf(−1.78,100000,0,1) and press ENTER.



TI-89

- In the Stats/List Editor, Press F5 (Distr) and choose Normal Cdf(.
- In the dialog box, enter these values: lower:−1.78, upper:100000, μ:0, σ:1, and then choose ENTER.

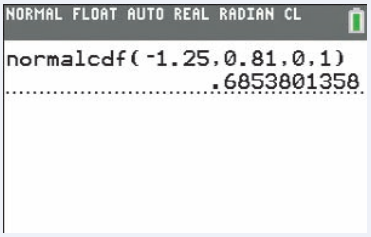


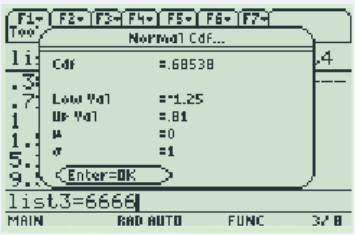
Note: We chose 100000 as the upper bound because it's many, many standard deviations above the mean.

These results agree with our previous answer using Table A: 0.9625.

2. What proportion of observations from the standard Normal distribution are between –1.25 and 0.81?

The screen shots below confirm our earlier result of 0.6854 using Table A.





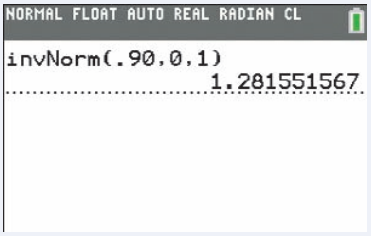
Working backward: The TI-83/84 and TI-89 invNorm function calculates the value corresponding to a given percentile in a Normal distribution. For this command, the syntax is invNorm(area to the left,mean,standard deviation).

3. What is the 90th percentile of the standard Normal distribution?

TI-83/84

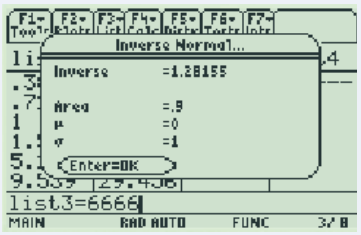
- Press 2nd VARS (Distr) and choose invNorm(. **OS 2.55 or later:** In the dialog box, enter these values: area:.90, μ:0, σ:1, choose Paste, and then press ENTER. **Older OS:** Complete the command invNorm(.90,0,1) and press ENTER.

These results match what we got using Table A.



TI-89

- In the Stats/List Editor, Press F5 (Distr), choose Inverse, and Inverse Normal….
- In the dialog box, enter these values: area:.90, μ:0, σ:1, and then choose ENTER.



Normal Distribution Calculations

We can answer a question about areas in *any* Normal distribution by standardizing and using Table A or by using technology. Here is an outline of the method for finding the proportion of the distribution in any region.

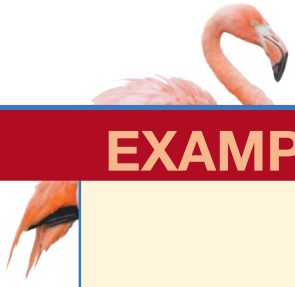
HOW TO FIND AREAS IN ANY NORMAL DISTRIBUTION

Step 1: State the distribution and the values of interest. Draw a Normal curve with the area of interest shaded and the mean, standard deviation, and boundary value(s) clearly identified.

Step 2: Perform calculations—show your work! Do one of the following: (i) Compute a z-score for each boundary value and use Table A or technology to find the desired area under the standard Normal curve; or (ii) use the normalcdf command and label each of the inputs.

Step 3: Answer the question.

Here's an example of the method at work.



EXAMPLE

Tiger on the Range

Normal calculations

On the driving range, Tiger Woods practices his swing with a particular club by hitting many, many balls. Suppose that when Tiger hits his driver, the distance the ball travels follows a Normal distribution with mean 304 yards and standard deviation 8 yards.

PROBLEM: What percent of Tiger's drives travel at least 290 yards?

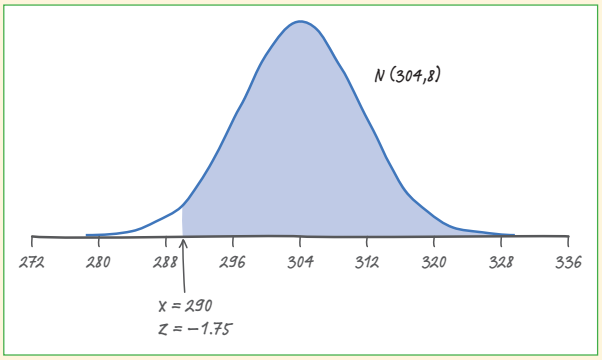


FIGURE 2.20 Distance traveled by Tiger Woods's drives on the range.

SOLUTION:

Step 1: State the distribution and the values of interest. The distance that Tiger's ball travels follows a Normal distribution with $\mu = 304$ and $\sigma = 8$. We want to find the percent of Tiger's drives that travel 290 yards or more. Figure 2.20 shows the distribution with the area of interest shaded and the mean, standard deviation, and boundary value labeled.

Step 2: Perform calculations—show your work! For the boundary value $x = 290$, we have

$$z = \frac{x - \mu}{\sigma} = \frac{290 - 304}{8} = -1.75$$

So drives of 290 yards or more correspond to $z \geq -1.75$ under the standard Normal curve.

From Table A, we see that the proportion of observations less than -1.75 is 0.0401. The area to the *right* of -1.75 is therefore $1 - 0.0401 = 0.9599$. This is about 0.96, or 96%.

Using technology: The command `normalcdf(lower:290,upper:100000,μ:304,σ:8)` also gives an area of 0.9599.

Step 3: Answer the question. About 96% of Tiger Woods's drives on the range travel at least 290 yards.

For Practice *Try Exercise* 53(a)

THINK ABOUT IT

What proportion of Tiger Woods's drives go exactly 290 yards? There is *no* area under the Normal density curve in Figure 2.20 exactly over the point 290. So the answer to our question based on the Normal model is 0. Tiger Woods's actual data may contain a drive that went exactly 290 yards (up to the precision of the measuring device). The Normal distribution is just an easy-to-use approximation, not a description of every detail in the data. One more thing: the areas under the curve with $x \geq 290$ and $x > 290$ are the same. According to the Normal model, the proportion of Tiger's drives that go at least 290 yards is the same as the proportion that go more than 290 yards.

The key to doing a Normal calculation is to sketch the area you want, then match that area with the area that the table (or technology) gives you. Here's another example.

EXAMPLE Tiger on the Range (Continued)

More complicated calculations

PROBLEM: What percent of Tiger's drives travel between 305 and 325 yards?

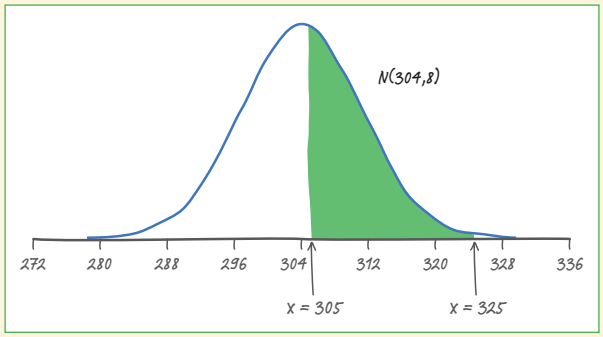


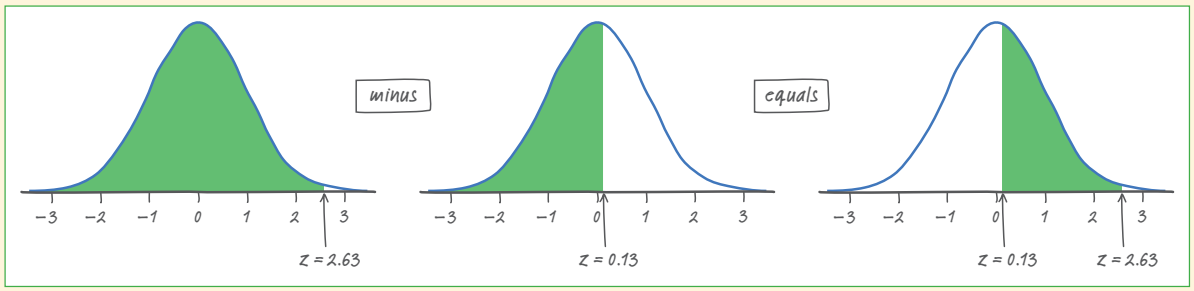
FIGURE 2.21 Distance traveled by Tiger Woods's drives on the range.

SOLUTION:

Step 1: State the distribution and the values of interest. As in the previous example, the distance that Tiger's ball travels follows a Normal distribution with $\mu = 304$ and $\sigma = 8$. We want to find the percent of Tiger's drives that travel between 305 and 325 yards. Figure 2.21 shows the distribution with the area of interest shaded and the mean, standard deviation, and boundary values labeled.

Step 2: Perform calculations—show your work! For the boundary value $x = 305$, $z = \dfrac{305 - 304}{8} = 0.13$. The standardized score for $x = 325$ is $z = \dfrac{325 - 304}{8} = 2.63$.

From Table A, we see that the area between $z = 0.13$ and $z = 2.63$ under the standard Normal curve is the area to the left of 2.63 minus the area to the left of 0.13. Look at the picture below to check this. From Table A, area between 0.13 and 2.63 = area to the left of 2.63 − area to the left of 0.13 = 0.9957 − 0.5517 = 0.4440.



Using technology: The command `normalcdf(lower:305,upper:325,μ:304,σ:8)` gives an area of 0.4459.

Step 3: Answer the question. About 45% of Tiger's drives travel between 305 and 325 yards.

For Practice *Try Exercise* 53(b)

Table A sometimes yields a slightly different answer from technology. That's because we have to round z-scores to two decimal places before using Table A.

Sometimes we encounter a value of z more extreme than those appearing in Table A. For example, the area to the left of $z = -4$ is not given directly in the table. The z-values in Table A leave only area 0.0002 in each tail unaccounted for. For practical purposes, we can act as if there is approximately zero area outside the range of Table A.

Working backwards: From areas to values: The previous two examples illustrated the use of Table A to find what proportion of the observations satisfies some condition, such as "Tiger's drive travels between 305 and 325 yards." Sometimes, we may want to find the observed value that corresponds to a given percentile. There are again three steps.

HOW TO FIND VALUES FROM AREAS IN ANY NORMAL DISTRIBUTION

Step 1: State the distribution and the values of interest. Draw a Normal curve with the area of interest shaded and the mean, standard deviation, and unknown boundary value clearly identified.

Step 2: Perform calculations—show your work! Do one of the following: (i) Use Table A or technology to find the value of z with the indicated area under the standard Normal curve, then "unstandardize" to transform back to the original distribution; or (ii) Use the invNorm command and label each of the inputs.

Step 3: Answer the question.

EXAMPLE

Cholesterol in Young Boys

Using Table A in reverse

High levels of cholesterol in the blood increase the risk of heart disease. For 14-year-old boys, the distribution of blood cholesterol is approximately Normal with mean $\mu = 170$ milligrams of cholesterol per deciliter of blood (mg/dl) and standard deviation $\sigma = 30$ mg/dl.[9]

PROBLEM: What is the 1st quartile of the distribution of blood cholesterol?

SOLUTION:

Step 1: State the distribution and the values of interest. The cholesterol level of 14-year-old boys follows a Normal distribution with $\mu = 170$ and $\sigma = 30$. The 1st quartile is the boundary value x with 25% of the distribution to its left. Figure 2.22 shows a picture of what we are trying to find.

Step 2: Perform calculations—show your work! Look in the body of Table A for the entry closest to 0.25. It's 0.2514. This is the entry corresponding to $z = -0.67$. So $z = -0.67$ is the standardized score with area 0.25 to its left. Now unstandardize. We know that the standardized score for the unknown cholesterol level x is $z = -0.67$. So x satisfies the equation $\dfrac{x - 170}{30} = -0.67$. Solving for x gives

$$x = 170 + (-0.67)(30) = 149.9$$

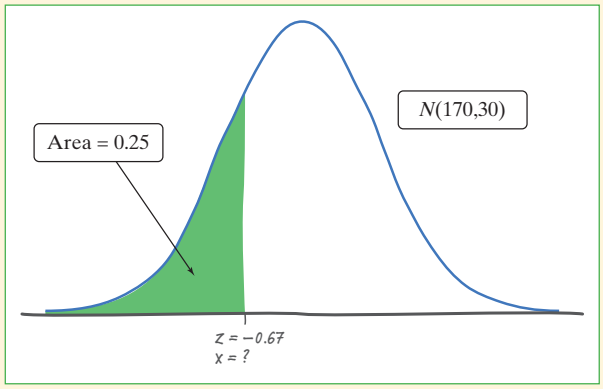


FIGURE 2.22 Locating the 1st quartile of the cholesterol distribution for 14-year-old boys.

Using technology: The command `invNorm(area:0.25,μ:170,σ:30)` gives $x = 149.8$.

Step 3: Answer the question. The 1st quartile of blood cholesterol levels in 14-year-old boys is about 150 mg/dl.

For Practice *Try Exercise* 53(c)

CHECK YOUR UNDERSTANDING

Follow the method shown in the examples to answer each of the following questions. Use your calculator or the *Normal Curve* applet to check your answers.

1. Cholesterol levels above 240 mg/dl may require medical attention. What percent of 14-year-old boys have more than 240 mg/dl of cholesterol?

2. People with cholesterol levels between 200 and 240 mg/dl are at considerable risk for heart disease. What percent of 14-year-old boys have blood cholesterol between 200 and 240 mg/dl?

3. What distance would a ball have to travel to be at the 80th percentile of Tiger Woods's drive lengths?

Assessing Normality

The Normal distributions provide good models for some distributions of real data. Examples include SAT and IQ test scores, the highway gas mileage of 2014 Corvette convertibles, state unemployment rates, and weights of 9-ounce bags of potato chips. The distributions of some other common variables are usually skewed and therefore distinctly non-Normal. Examples include economic variables such as personal income and total sales of business firms, the survival times of cancer patients after treatment, and the lifetime of electronic devices. While experience can suggest whether or not a Normal distribution is a reasonable model in a particular case, it is risky to assume that a distribution is Normal without actually inspecting the data.

In the latter part of this course, we will use various statistical inference procedures to try to answer questions that are important to us. These tests involve sampling individuals and recording data to gain insights about the populations from which they come. Many of these procedures are based on the assumption that the population is approximately Normally distributed. Consequently, we need to develop a strategy for assessing Normality.

EXAMPLE Unemployment in the States

Are the data close to Normal?

Let's start by examining data on unemployment rates in the 50 states. Here are the data arranged from lowest (North Dakota's 4.1%) to highest (Michigan's 14.7%).[10]

4.1	4.5	5.0	6.3	6.3	6.4	6.4	6.6	6.7	6.7	6.7	6.9	7.0
7.0	7.2	7.4	7.4	7.4	7.8	8.0	8.0	8.2	8.2	8.4	8.5	8.5
8.6	8.7	8.8	8.9	9.1	9.2	9.5	9.6	9.6	9.7	10.2	10.3	10.5
10.6	10.6	10.8	10.9	11.1	11.5	12.3	12.3	12.3	12.7	14.7		

• *Plot the data.* Make a dotplot, stemplot, or histogram. See if the graph is approximately symmetric and bell-shaped.

Figure 2.23 is a histogram of the state unemployment rates. The graph is roughly symmetric, single-peaked, and somewhat bell-shaped.

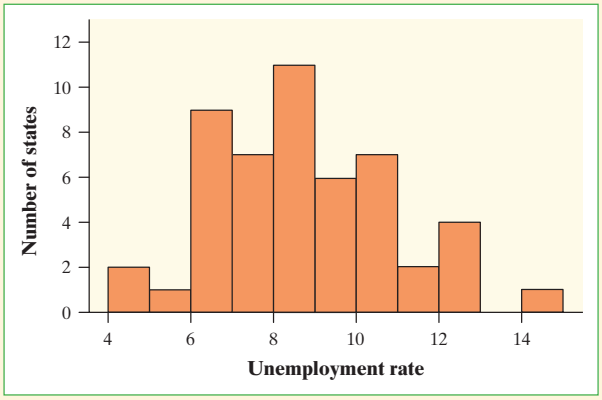


FIGURE 2.23 Histogram of state unemployment rates.

• *Check whether the data follow the 68–95–99.7 rule.*

We entered the unemployment rates into computer software and requested summary statistics. Here's what we got:

```
Mean = 8.682 Standard deviation = 2.225.
```

Now we can count the number of observations within one, two, and three standard deviations of the mean.

Mean ± 1 SD:	6.457 to 10.907	36 out of 50 = 72%
Mean ± 2 SD:	4.232 to 13.132	48 out of 50 = 96%
Mean ± 3 SD:	2.007 to 15.357	50 out of 50 = 100%

These percents are quite close to the 68%, 95%, and 99.7% targets for a Normal distribution.

If a graph of the data is clearly skewed, has multiple peaks, or isn't bell-shaped, that's evidence that the distribution is *not* Normal. However, just because a plot of the data *looks* Normal, we can't say that the distribution *is* Normal. The 68–95–99.7 rule can give additional evidence in favor of or against Normality. A **Normal probability plot** also provides a good assessment of whether a data set follows a Normal distribution.

EXAMPLE

Unemployment in the States

Making a Normal probability plot

Most software packages, including Minitab, Fathom, and JMP, can construct Normal probability plots (sometimes called Normal quantile plots) from entered data. The TI-83/84 and TI-89 will also make these graphs. Here's how a Normal probability plot is constructed.

1. *Arrange the observed data values from smallest to largest.* Record the percentile corresponding to each observation (but remember that there are several definitions of "percentile"). For example, the smallest observation in a set of 50 values is at either the 0th percentile (because 0 out of 50 values are below this observation) or the 2nd percentile (because 1 out of 50 values are at or below this observation). Technology usually "splits the difference," declaring this minimum value to be at the (0 + 2)/2 = 1st percentile. By similar reasoning, the second-smallest value is at the 3rd percentile, the third-smallest value is at the 5th percentile, and so on. The maximum value is at the (98 + 100)/2 = 99th percentile.

2. *Use the standard Normal distribution (Table A or* `invNorm`*) to find the z-scores at these same percentiles.* For example, the 1st percentile of the standard Normal distribution is $z = -2.326$. The 3rd percentile is $z = -1.881$; the 5th percentile is $z = -1.645$; . . . ; the 99th percentile is $z = 2.326$.

3. *Plot each observation x against its expected z-score from Step 2.* If the data distribution is close to Normal, the plotted points will lie close to some straight line. Figure 2.24 shows a Normal probability plot for the state unemployment data. There is a strong linear pattern, which suggests that the distribution of unemployment rates is close to Normal.

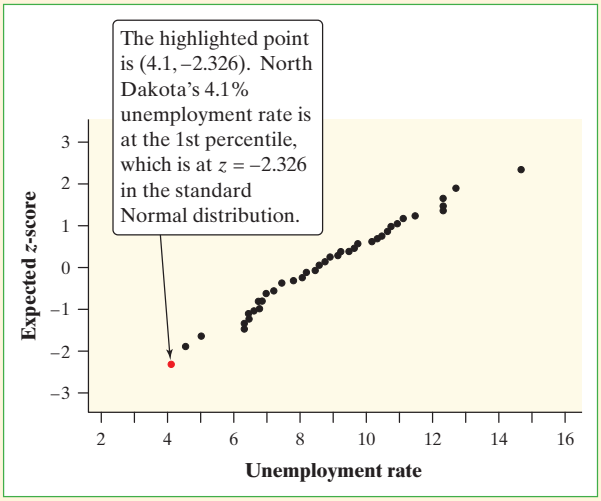


FIGURE 2.24 Normal probability plot of the percent of unemployed individuals in each of the 50 states.

Some software plots the data values on the horizontal axis and the *z*-scores on the vertical axis, while other software does just the reverse. The TI-83/84 and TI-89 give you both options. We prefer the data values on the horizontal axis, which is consistent with other types of graphs we have made.

As Figure 2.24 indicates, real data almost always show some departure from Normality. *When you examine a Normal probability plot, look for shapes that show clear departures from Normality. Don't overreact to minor wiggles in the plot.* When we discuss statistical methods that are based on the Normal model, we will pay attention to the sensitivity of each method to departures from Normality. Many common methods work well as long as the data are approximately Normal.

INTERPRETING NORMAL PROBABILITY PLOTS

If the points on a **Normal probability plot** lie close to a straight line, the data are approximately Normally distributed. Systematic deviations from a straight line indicate a non-Normal distribution. Outliers appear as points that are far away from the overall pattern of the plot.

AP® EXAM TIP Normal probability plots are not included on the AP® Statistics topic outline. However, these graphs are very useful for assessing Normality. You may use them on the AP® exam if you wish—just be sure that you know what you're looking for (a linear pattern).

Let's look at an example of some data that are *not* Normally distributed.

EXAMPLE Guinea Pig Survival

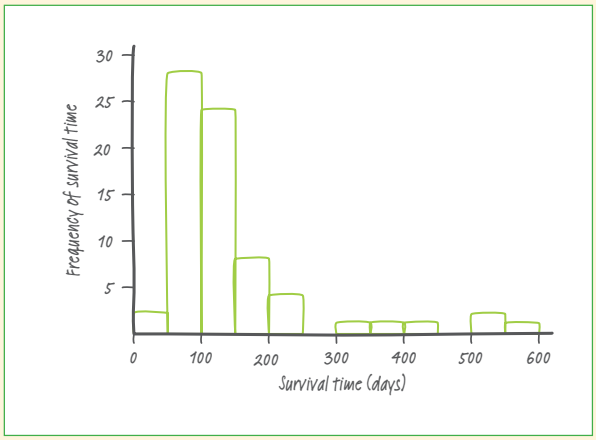
Assessing Normality

In Chapter 1 Review Exercise R1.7 (page 77), we introduced data on the survival times in days of 72 guinea pigs after they were injected with infectious bacteria in a medical experiment.

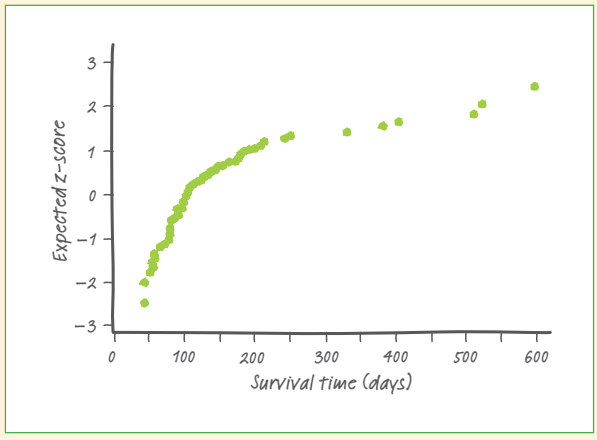
PROBLEM: Determine whether these data are approximately Normally distributed.

SOLUTION: Let's follow the first step in our strategy for assessing Normality: plot the data! Figure 2.25(a) shows a histogram of the guinea pig survival times. We can see that the distribution is heavily right-skewed. Figure 2.25(b) is a Normal probability plot of the data. The clear curvature in this graph confirms that these data do not follow a Normal distribution.

We won't bother checking the 68–95–99.7 rule for these data because the graphs in Figure 2.25 indicate serious departures from Normality.



(a)



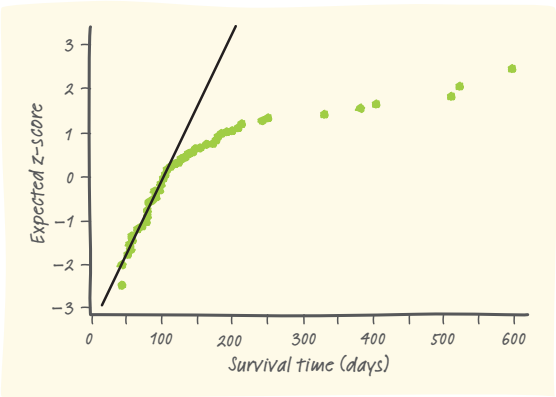
(b)

FIGURE 2.25 (a) Histogram and (b) Normal probability plot of the guinea pig survival data.

For Practice *Try Exercise* 63

THINK ABOUT IT

How can we determine shape from a Normal probability plot? Look at the Normal probability plot of the guinea pig survival data in Figure 2.25(b). Imagine drawing a line through the leftmost points, which correspond to the smaller observations. The larger observations fall systematically



to the right of this line. That is, the right-of-center observations have much larger values than expected based on their percentiles and the corresponding z-scores from the standard Normal distribution.

This Normal probability plot indicates that the guinea pig survival data are strongly right-skewed. *In a right-skewed distribution, the largest observations fall distinctly to the right of a line drawn through the main body of points.* Similarly, left skewness is evident when the smallest observations fall to the left the line.

If you're wondering how to make a Normal probability plot on your calculator, the following Technology Corner shows you the process.

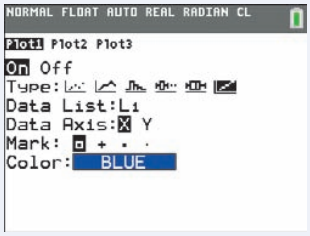
6. TECHNOLOGY CORNER NORMAL PROBABILITY PLOTS

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

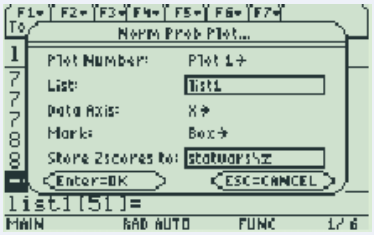
To make a Normal probability plot for a set of quantitative data:

- Enter the data values in L1/list1. We'll use the state unemployment rates data from page 122.
- Define Plot1 as shown.

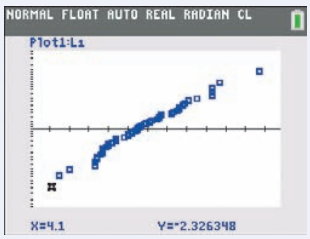
TI-83/84

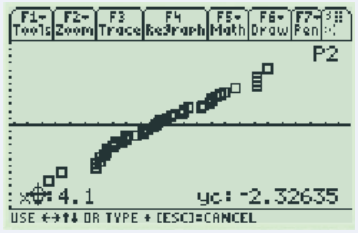


TI-89



- Use `ZoomStat` (`ZoomData` on the TI-89) to see the finished graph.





Interpretation: The Normal probability plot is quite linear, so it is reasonable to believe that the data follow a Normal distribution.

DATA EXPLORATION The vending machine problem

Have you ever purchased a hot drink from a vending machine? The intended sequence of events runs something like this. You insert your money into the machine and select your preferred beverage. A cup falls out of the machine, landing upright. Liquid pours out until the cup is nearly full. You reach in, grab the piping-hot cup, and drink happily.

Sometimes, things go wrong. The machine might swipe your money. Or the cup might fall over. More frequently, everything goes smoothly until the liquid begins to flow. It might stop flowing when the cup is only half full. Or the liquid might keep coming until your cup overflows. Neither of these results leaves you satisfied.

The vending machine company wants to keep customers happy. So they have decided to hire you as a statistical consultant. They provide you with the following summary of important facts about the vending machine:

- Cups will hold 8 ounces.
- The amount of liquid dispensed varies according to a Normal distribution centered at the mean μ that is set in the machine.
- $\sigma = 0.2$ ounces.

If a cup contains too much liquid, a customer may get burned from a spill. This could result in an expensive lawsuit for the company. On the other hand, customers may be irritated if they get a cup with too little liquid from the machine. Given these issues, what mean setting for the machine would you recommend? Write a brief report to the vending machine company president that explains your answer.

case closed

Do You Sudoku?

In the chapter-opening Case Study (page 83), one of the authors played an online game of sudoku. At the end of his game, the graph on the next page was displayed. The density curve shown was constructed from a histogram of times from 4,000,000 games played in one week at this Web site. You will now use what you have learned in this chapter to analyze how well the author did.

1. State and interpret the percentile for the author's time of 3 minutes and 19 seconds. (Remember that smaller times indicate better performance.)
2. Explain why you cannot find the z-score corresponding to the author's time.
3. Suppose the author's time to finish the puzzle had been 5 minutes and 6 seconds instead.
   (a) Would his percentile be greater than 50%, equal to 50%, or less than 50%? Justify your answer.
   (b) Would his z-score be positive, negative, or zero? Explain.

4. From long experience, the author's times to finish an easy sudoku puzzle at this Web site follow a Normal distribution with mean 4.2 minutes and standard deviation 0.7 minutes. In what percent of the games that he plays does the author finish an easy puzzle in less than 3 minutes and 15 seconds? Show your work. (*Note*: 3 minutes and 15 seconds is not the same as 3.15 seconds!)
5. The author's wife also enjoys playing sudoku online. Her times to finish an easy puzzle at this Web site follow a Normal distribution with mean 3.8 minutes and standard deviation 0.9 minutes. In her most recent game, she finished in 3 minutes. Whose performance is better, relatively speaking: the author's 3 minutes and 19 seconds or his wife's 3 minutes? Justify your answer.

Section 2.2 Summary

- We can describe the overall pattern of a distribution by a **density curve**. A density curve always remains on or above the horizontal axis and has total area 1 underneath it. An area under a density curve gives the proportion of observations that fall in an interval of values.
- A density curve is an idealized description of the overall pattern of a distribution that smooths out the irregularities in the actual data. We write the **mean of a density curve** as μ and the **standard deviation of a density curve** as σ to distinguish them from the mean $\bar{x}$ and the standard deviation s_x of the actual data.
- The mean and the median of a density curve can be located by eye. The mean μ is the balance point of the curve. The median divides the area under the curve in half. The standard deviation σ cannot be located by eye on most density curves.
- The mean and median are equal for symmetric density curves. The mean of a skewed curve is located farther toward the long tail than the median is.
- The **Normal distributions** are described by a special family of bell-shaped, symmetric density curves, called **Normal curves**. The mean μ and standard deviation σ completely specify a Normal distribution $N(\mu,\sigma)$. The mean is the center of the curve, and σ is the distance from μ to the change-of-curvature points on either side.
- All Normal distributions obey the **68–95–99.7 rule,** which describes what percent of observations lie within one, two, and three standard deviations of the mean.
- All Normal distributions are the same when measurements are standardized. If x follows a Normal distribution with mean μ and standard deviation σ, we can standardize using

$$z = \frac{x - \mu}{\sigma}$$

The variable z has the **standard Normal distribution** with mean 0 and standard deviation 1.

- **Table A** at the back of the book gives percentiles for the standard Normal curve. By standardizing, we can use Table A to determine the percentile for a given z-score or the z-score corresponding to a given percentile in any Normal distribution. You can use your calculator or the *Normal Curve* applet to perform Normal calculations quickly.
- To perform certain inference procedures in later chapters, we will need to know that the data come from populations that are approximately Normally distributed. To assess Normality for a given set of data, we first observe the shape of a dotplot, stemplot, or histogram. Then we can check how well the data fit the 68–95–99.7 rule for Normal distributions. Another good method for assessing Normality is to construct a **Normal probability plot**.

2.2 TECHNOLOGY CORNERS

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

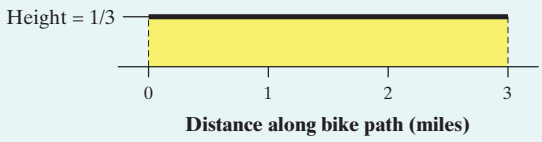
Section 2.2 Exercises

33. **Density curves** Sketch a density curve that might describe a distribution that is symmetric but has two peaks.

34. **Density curves** Sketch a density curve that might describe a distribution that has a single peak and is skewed to the left.

Exercises 35 to 38 involve a special type of density curve—one that takes constant height (looks like a horizontal line) over some interval of values. This density curve describes a variable whose values are distributed evenly (uniformly) over some interval of values. We say that such a variable has a ***uniform distribution****.*

35. **Biking accidents** Accidents on a level, 3-mile bike path occur uniformly along the length of the path. The figure below displays the density curve that describes the uniform distribution of accidents.

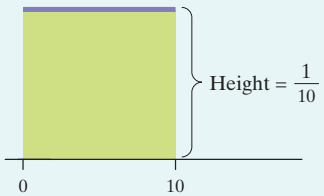


(a) Explain why this curve satisfies the two requirements for a density curve.

(b) The proportion of accidents that occur in the first mile of the path is the area under the density curve between 0 miles and 1 mile. What is this area?

(c) Sue's property adjoins the bike path between the 0.8 mile mark and the 1.1 mile mark. What proportion of accidents happen in front of Sue's property? Explain.

36. **Where's the bus?** Sally takes the same bus to work every morning. The amount of time (in minutes) that she has to wait for the bus to arrive is described by the uniform distribution below.



(a) Explain why this curve satisfies the two requirements for a density curve.

(b) On what percent of days does Sally have to wait more than 8 minutes for the bus?

(c) On what percent of days does Sally wait between 2.5 and 5.3 minutes for the bus?

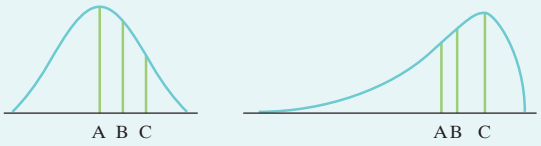
37. **Biking accidents** What is the mean μ of the density curve pictured in Exercise 35? (That is, where would the curve balance?) What is the median? (That is, where is the point with area 0.5 on either side?)

38. **Where's the bus?** What is the mean μ of the density curve pictured in Exercise 36? What is the median?

39. **Mean and median** The figure below displays two density curves, each with three points marked. At which of these points on each curve do the mean and the median fall?

(a) (b)

40. **Mean and median** The figure below displays two density curves, each with three points marked. At which of these points on each curve do the mean and the median fall?



(a) (b)

41. **Men's heights** The distribution of heights of adult American men is approximately Normal with mean 69 inches and standard deviation 2.5 inches. Draw an accurate sketch of the distribution of men's heights. Be sure to label the mean, as well as the points 1, 2, and 3 standard deviations away from the mean on the horizontal axis.

42. **Potato chips** The distribution of weights of 9-ounce bags of a particular brand of potato chips is approximately Normal with mean $\mu = 9.12$ ounces and standard deviation $\sigma = 0.05$ ounce. Draw an accurate sketch of the distribution of potato chip bag weights. Be sure to label the mean, as well as the points 1, 2, and 3 standard deviations away from the mean on the horizontal axis.

43. **Men's heights** Refer to Exercise 41. Use the 68–95–99.7 rule to answer the following questions. Show your work!

pg 111

(a) Between what heights do the middle 95% of men fall?

(b) What percent of men are taller than 74 inches?

(c) What percent of men are between 64 and 66.5 inches tall?

(d) A height of 71.5 inches corresponds to what percentile of adult male American heights?

44. **Potato chips** Refer to Exercise 42. Use the 68–95–99.7 rule to answer the following questions. Show your work!

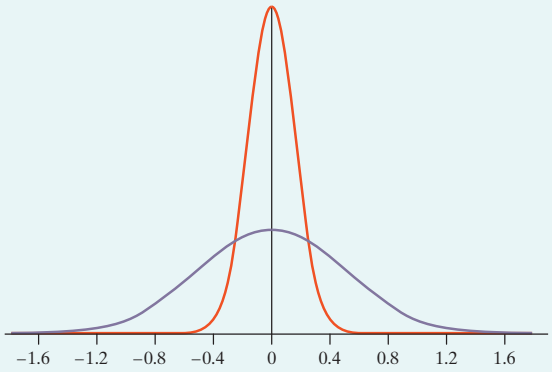
(a) Between what weights do the middle 68% of bags fall?

(b) What percent of bags weigh less than 9.02 ounces?

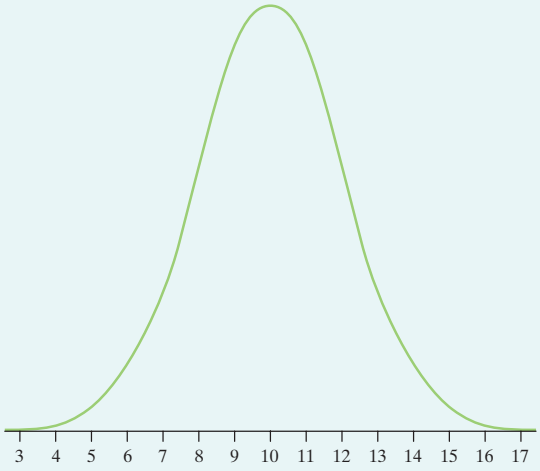
(c) What percent of 9-ounce bags of this brand of potato chips weigh between 8.97 and 9.17 ounces?

(d) A bag that weighs 9.07 ounces is at what percentile in this distribution?

45. **Estimating SD** The figure below shows two Normal curves, both with mean 0. Approximately what is the standard deviation of each of these curves?



46. **A Normal curve** Estimate the mean and standard deviation of the Normal density curve in the figure below.



For Exercises 47 to 50, use Table A to find the proportion of observations from the standard Normal distribution that satisfies each of the following statements. In each case, sketch a standard Normal curve and shade the area under the curve that is the answer to the question.

47. **Table A practice**

(a) $z < 2.85$ (c) $z > -1.66$

(b) $z > 2.85$ (d) $-1.66 < z < 2.85$

48. Table A practice

(a) $z < -2.46$ (c) $0.89 < z < 2.46$

(b) $z > 2.46$ (d) $-2.95 < z < -1.27$

pg 115 **49. More Table A practice**

(a) z is between -1.33 and 1.65

(b) z is between 0.50 and 1.79

50. More Table A practice

(a) z is between -2.05 and 0.78

(b) z is between -1.11 and -0.32

For Exercises 51 and 52, use Table A to find the value z from the standard Normal distribution that satisfies each of the following conditions. In each case, sketch a standard Normal curve with your value of z marked on the axis.

51. Working backward

(a) The 10th percentile.

(b) 34% of all observations are greater than z.

52. Working backward

(a) The 63rd percentile.

(b) 75% of all observations are greater than z.

53. Length of pregnancies The length of human pregnancies from conception to birth varies according to a distribution that is approximately Normal with mean 266 days and standard deviation 16 days.

pg 118 (a) At what percentile is a pregnancy that lasts 240 days (that's about 8 months)?

pg 119 (b) What percent of pregnancies last between 240 and 270 days (roughly between 8 months and 9 months)?

pg 120 (c) How long do the longest 20% of pregnancies last?

54. IQ test scores Scores on the Wechsler Adult Intelligence Scale (a standard IQ test) for the 20 to 34 age group are approximately Normally distributed with $\mu = 110$ and $\sigma = 25$.

(a) At what percentile is an IQ score of 150?

(b) What percent of people aged 20 to 34 have IQs between 125 and 150?

(c) MENSA is an elite organization that admits as members people who score in the top 2% on IQ tests. What score on the Wechsler Adult Intelligence Scale would an individual aged 20 to 34 have to earn to qualify for MENSA membership?

55. Put a lid on it! At some fast-food restaurants, customers who want a lid for their drinks get them from a large stack left near straws, napkins, and condiments. The lids are made with a small amount of flexibility so they can be stretched across the mouth of the cup and then snugly secured. When lids are too small or too large, customers can get very frustrated, especially if they end up spilling their drinks. At one particular restaurant, large drink cups require lids with a "diameter" of between 3.95 and 4.05 inches. The restaurant's lid supplier claims that the diameter of their large lids follows a Normal distribution with mean 3.98 inches and standard deviation 0.02 inches. Assume that the supplier's claim is true.

(a) What percent of large lids are too small to fit? Show your method.

(b) What percent of large lids are too big to fit? Show your method.

(c) Compare your answers to parts (a) and (b). Does it make sense for the lid manufacturer to try to make one of these values larger than the other? Why or why not?

56. I think I can! An important measure of the performance of a locomotive is its "adhesion," which is the locomotive's pulling force as a multiple of its weight. The adhesion of one 4400-horsepower diesel locomotive varies in actual use according to a Normal distribution with mean $\mu = 0.37$ and standard deviation $\sigma = 0.04$.

(a) For a certain small train's daily route, the locomotive needs to have an adhesion of at least 0.30 for the train to arrive at its destination on time. On what proportion of days will this happen? Show your method.

(b) An adhesion greater than 0.50 for the locomotive will result in a problem because the train will arrive too early at a switch point along the route. On what proportion of days will this happen? Show your method.

(c) Compare your answers to parts (a) and (b). Does it make sense to try to make one of these values larger than the other? Why or why not?

57. Put a lid on it! Refer to Exercise 55. The supplier is considering two changes to reduce the percent of its large-cup lids that are too small to 1%. One strategy is to adjust the mean diameter of its lids. Another option is to alter the production process, thereby decreasing the standard deviation of the lid diameters.

(a) If the standard deviation remains at $\sigma = 0.02$ inches, at what value should the supplier set the mean diameter of its large-cup lids so that only 1% are too small to fit? Show your method.

(b) If the mean diameter stays at $\mu = 3.98$ inches, what value of the standard deviation will result in only 1% of lids that are too small to fit? Show your method.

(c) Which of the two options in parts (a) and (b) do you think is preferable? Justify your answer. (Be sure to consider the effect of these changes on the percent of lids that are too large to fit.)

58. **I think I can!** Refer to Exercise 56. The locomotive's manufacturer is considering two changes that could reduce the percent of times that the train arrives late. One option is to increase the mean adhesion of the locomotive. The other possibility is to decrease the variability in adhesion from trip to trip, that is, to reduce the standard deviation.

(a) If the standard deviation remains at $\sigma = 0.04$, to what value must the manufacturer change the mean adhesion of the locomotive to reduce its proportion of late arrivals to only 2% of days? Show your work.

(b) If the mean adhesion stays at $\mu = 0.37$, how much must the standard deviation be decreased to ensure that the train will arrive late only 2% of the time? Show your work.

(c) Which of the two options in parts (a) and (b) do you think is preferable? Justify your answer. (Be sure to consider the effect of these changes on the percent of days that the train arrives early to the switch point.)

59. **Deciles** The deciles of any distribution are the values at the 10th, 20th, . . . , 90th percentiles. The first and last deciles are the 10th and the 90th percentiles, respectively.

(a) What are the first and last deciles of the standard Normal distribution?

(b) The heights of young women are approximately Normal with mean 64.5 inches and standard deviation 2.5 inches. What are the first and last deciles of this distribution? Show your work.

60. **Outliers** The percent of the observations that are classified as outliers by the $1.5 \times IQR$ rule is the same in any Normal distribution. What is this percent? Show your method clearly.

61. **Flight times** An airline flies the same route at the same time each day. The flight time varies according to a Normal distribution with unknown mean and standard deviation. On 15% of days, the flight takes more than an hour. On 3% of days, the flight lasts 75 minutes or more. Use this information to determine the mean and standard deviation of the flight time distribution.

62. **Brush your teeth** The amount of time Ricardo spends brushing his teeth follows a Normal distribution with unknown mean and standard deviation. Ricardo spends less than one minute brushing his teeth about 40% of the time. He spends more than two minutes brushing his teeth 2% of the time. Use this information to determine the mean and standard deviation of this distribution.

63. **Sharks** Here are the lengths in feet of 44 great white sharks:[11]

pg 124

18.7	12.3	18.6	16.4	15.7	18.3	14.6	15.8	14.9	17.6	12.1
16.4	16.7	17.8	16.2	12.6	17.8	13.8	12.2	15.2	14.7	12.4
13.2	15.8	14.3	16.6	9.4	18.2	13.2	13.6	15.3	16.1	13.5
19.1	16.2	22.8	16.8	13.6	13.2	15.7	19.7	18.7	13.2	16.8

(a) Enter these data into your calculator and make a histogram. Include a sketch of the graph on your paper. Then calculate one-variable statistics. Describe the shape, center, and spread of the distribution of shark lengths.

(b) Calculate the percent of observations that fall within 1, 2, and 3 standard deviations of the mean. How do these results compare with the 68–95–99.7 rule?

(c) Use your calculator to construct a Normal probability plot. Include a sketch of the graph on your paper. Interpret this plot.

(d) Having inspected the data from several different perspectives, do you think these data are approximately Normal? Write a brief summary of your assessment that combines your findings from parts (a) through (c).

64. **Density of the earth** In 1798, the English scientist Henry Cavendish measured the density of the earth several times by careful work with a torsion balance. The variable recorded was the density of the earth as a multiple of the density of water. Here are Cavendish's 29 measurements:[12]

5.50	5.61	4.88	5.07	5.26	5.55	5.36	5.29	5.58	5.65
5.57	5.53	5.62	5.29	5.44	5.34	5.79	5.10	5.27	5.39
5.42	5.47	5.63	5.34	5.46	5.30	5.75	5.68	5.85	

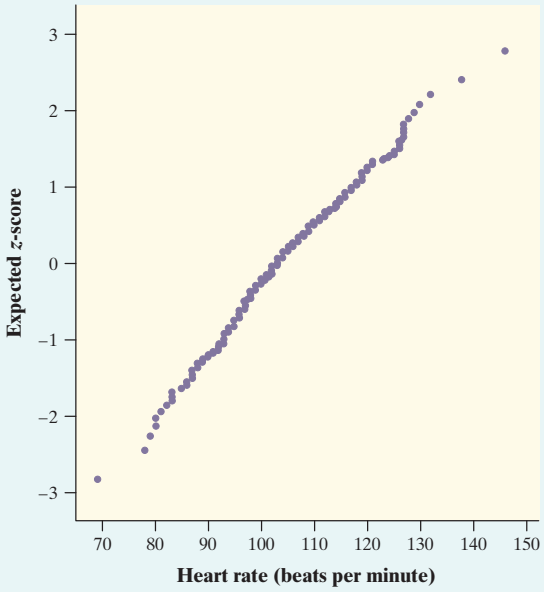
(a) Enter these data into your calculator and make a histogram. Include a sketch of the graph on your paper. Then calculate one-variable statistics. Describe the shape, center, and spread of the distribution of density measurements.

(b) Calculate the percent of observations that fall within 1, 2, and 3 standard deviations of the mean. How do these results compare with the 68–95–99.7 rule?

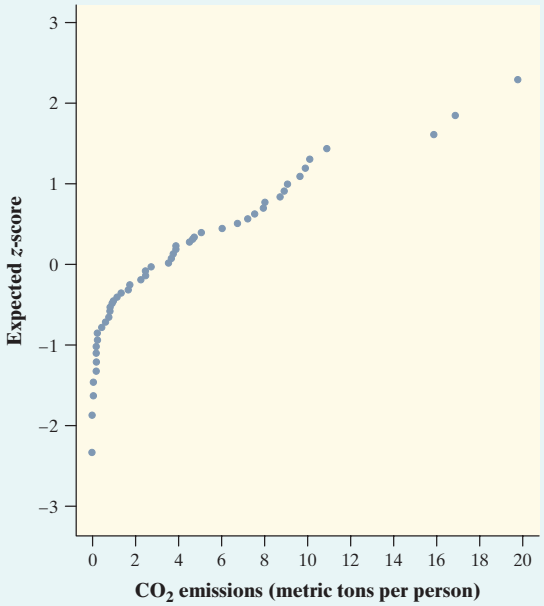
(c) Use your calculator to construct a Normal probability plot. Include a sketch of the graph on your paper. Interpret this plot.

(d) Having inspected the data from several different perspectives, do you think these data are approximately Normal? Write a brief summary of your assessment that combines your findings from parts (a) through (c).

65. **Runners' heart rates** The figure below is a Normal probability plot of the heart rates of 200 male runners after six minutes of exercise on a treadmill.[13] The distribution is close to Normal. How can you see this? Describe the nature of the small deviations from Normality that are visible in the plot.



66. **Carbon dioxide emissions** The figure below is a Normal probability plot of the emissions of carbon dioxide per person in 48 countries.[14] In what ways is this distribution non-Normal?



67. **Is Michigan Normal?** We collected data on the tuition charged by colleges and universities in Michigan. Here are some numerical summaries for the data:

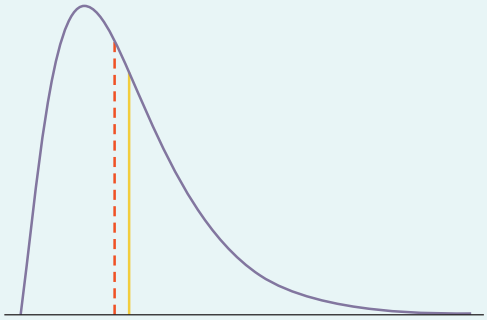
Mean	Std. Dev.	Min	Max
10614	8049	1873	30823

Based on the relationship between the mean, standard deviation, minimum, and maximum, is it reasonable to believe that the distribution of Michigan tuitions is approximately Normal? Explain.

68. **Weights aren't Normal** The heights of people of the same gender and similar ages follow Normal distributions reasonably closely. Weights, on the other hand, are not Normally distributed. The weights of women aged 20 to 29 have mean 141.7 pounds and median 133.2 pounds. The first and third quartiles are 118.3 pounds and 157.3 pounds. What can you say about the shape of the weight distribution? Why?

Multiple choice: Select the best answer for Exercises 69 to 74.

69. Two measures of center are marked on the density curve shown. Which of the following is correct?



(a) The median is at the yellow line and the mean is at the red line.

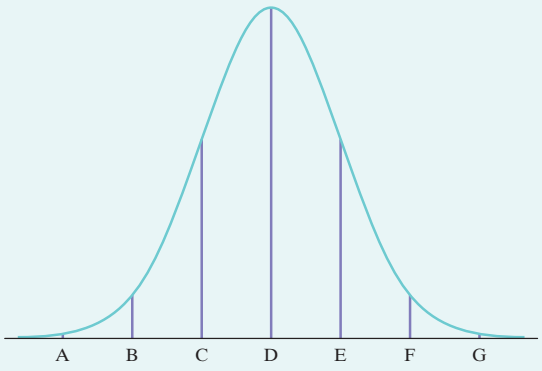
(b) The median is at the red line and the mean is at the yellow line.

(c) The mode is at the red line and the median is at the yellow line.

(d) The mode is at the yellow line and the median is at the red line.

(e) The mode is at the red line and the mean is at the yellow line.

Exercises 70 to 72 refer to the following setting. The weights of laboratory cockroaches follow a Normal distribution with mean 80 grams and standard deviation 2 grams. The following figure is the Normal curve for this distribution of weights.



70. Point C on this Normal curve corresponds to

(a) 84 grams. (c) 78 grams. (e) 74 grams.
(b) 82 grams. (d) 76 grams.

71. About what percent of the cockroaches have weights between 76 and 84 grams?

(a) 99.7% (c) 68% (e) 34%
(b) 95% (d) 47.5%

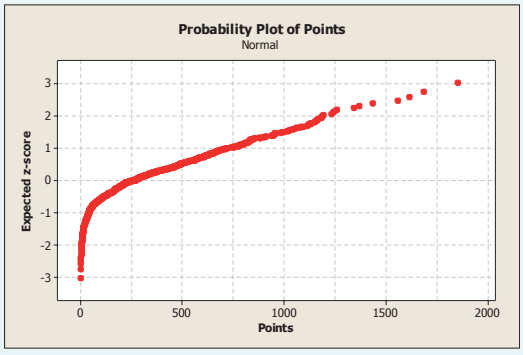
72. About what proportion of the cockroaches will have weights greater than 83 grams?

(a) 0.0228 (c) 0.1587 (e) 0.0772
(b) 0.0668 (d) 0.9332

73. A different species of cockroach has weights that follow a Normal distribution with a mean of 50 grams. After measuring the weights of many of these cockroaches, a lab assistant reports that 14% of the cockroaches weigh more than 55 grams. Based on this report, what is the approximate standard deviation of weights for this species of cockroaches?

(a) 4.6 (d) 14.0
(b) 5.0 (e) Cannot determine without more information.
(c) 6.2

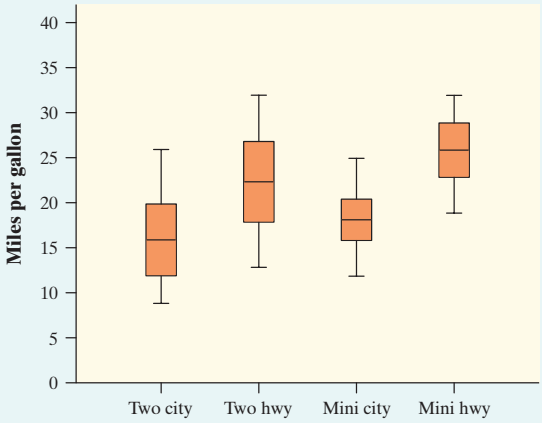
74. The following Normal probability plot shows the distribution of points scored for the 551 players in the 2011–2012 NBA season.



If the distribution of points was displayed in a histogram, what would be the best description of the histogram's shape?

(a) Approximately Normal
(b) Symmetric but not approximately Normal
(c) Skewed left
(d) Skewed right
(e) Cannot be determined

75. Gas it up! (1.3) Interested in a sporty car? Worried that it might use too much gas? The Environmental Protection Agency lists most such vehicles in its "two-seater" or "minicompact" categories. The figure shows boxplots for both city and highway gas mileages for our two groups of cars. Write a few sentences comparing these distributions.



76. Python eggs (1.1) How is the hatching of water python eggs influenced by the temperature of the snake's nest? Researchers assigned newly laid eggs to one of three temperatures: hot, neutral, or cold. Hot duplicates the extra warmth provided by the mother python, and cold duplicates the absence of the mother. Here are the data on the number of eggs and the number that hatched:[15]

	Cold	Neutral	Hot
Number of eggs	27	56	104
Number hatched	16	38	75

(a) Make a two-way table of temperature by outcome (hatched or not).

(b) Calculate the percent of eggs in each group that hatched. The researchers believed that eggs would be less likely to hatch in cold water. Do the data support that belief?

FRAPPY! Free Response AP® Problem, Yay!

The following problem is modeled after actual AP® Statistics exam free response questions. Your task is to generate a complete, concise response in 15 minutes.

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

The distribution of scores on a recent test closely followed a Normal distribution with a mean of 22 points and a standard deviation of 4 points.

(a) What proportion of the students scored at least 25 points on this test?

(b) What is the 31st percentile of the distribution of test scores?

(c) The teacher wants to transform the test scores so that they have an approximately Normal distribution with a mean of 80 points and a standard deviation of 10 points. To do this, she will use a formula in the form:

$$new\ score = a + b\ (old\ score)$$

Find the values of a and b that the teacher should use to transform the distribution of test scores.

(d) Before the test, the teacher gave a review assignment for homework. The maximum score on the assignment was 10 points. The distribution of scores on this assignment had a mean of 9.2 points and a standard deviation of 2.1 points. Would it be appropriate to use a Normal distribution to calculate the proportion of students who scored below 7 points on this assignment? Explain.

After you finish, you can view two example solutions on the book's Web site (www.whfreeman.com/tps5e). Determine whether you think each solution is "complete," "substantial," "developing," or "minimal." If the solution is not complete, what improvements would you suggest to the student who wrote it? Finally, your teacher will provide you with a scoring rubric. Score your response and note what, if anything, you would do differently to improve your own score.

Chapter Review

Section 2.1: Describing Location in a Distribution

In this section, you learned two different ways to describe the location of individuals in a distribution, percentiles and standardized scores (z-scores). Percentiles describe the location of an individual by measuring what percent of the observations in the distribution have a value less than the individual's value. A cumulative relative frequency graph is a handy tool for identifying percentiles in a distribution. You can use it to estimate the percentile for a particular value of a variable or estimate the value of the variable at a particular percentile.

Standardized scores (z-scores) describe the location of an individual in a distribution by measuring how many standard deviations the individual is above or below the mean. To find the standardized score for a particular observation, transform the value by subtracting the mean and dividing the difference by the standard deviation. Besides describing the location of an individual in a distribution, you can also use z-scores to compare observations from different distributions—standardizing the values puts them on a standard scale.

You also learned to describe the effects on the shape, center, and spread of a distribution when transforming data from one scale to another. Adding a positive constant to (or subtracting it from) each value in a data set changes the measures of location but not the shape or spread of the distribution. Multiplying or dividing each value in a data set by a positive constant changes the measures of location and measures of spread but not the shape of the distribution.

Section 2.2: Density Curves and Normal Distributions

In this section, you learned how density curves are used to model distributions of data. An area under a density curve gives the proportion of observations that fall in a specified interval of values. The total area under a density curve is 1, or 100%.

The most commonly used density curve is called a Normal curve. The Normal curve is symmetric, single-peaked, and bell-shaped with mean μ and standard deviation σ. For any distribution of data that is approximately Normal in shape, about 68% of the observations will be within 1 standard deviation of the mean, about 95% of the observations will be within 2 standard deviations of the mean, and about 99.7% of the observations will be within 3 standard deviations of the mean. Conveniently, this relationship is called the 68–95–99.7 rule.

When observations do not fall exactly 1, 2, or 3 standard deviations from the mean, you learned how to use Table A (or technology) to identify the proportion of values in any specified interval under a Normal curve. You also learned how to use Table A (or technology) to determine the value of an individual that falls at a specified percentile in a Normal distribution. On the AP® exam, it is extremely important that you clearly communicate your methods when answering questions that involve the Normal distribution. You must specify the shape (Normal), center (μ), and spread (σ) of the distribution; identify the region under the Normal curve that you are working with; and correctly calculate the answer with work shown. Shading a Normal curve with the mean, standard deviation, and boundaries clearly identified is a great start.

Finally, you learned how to determine whether a distribution of data is approximately Normal using graphs (dotplots, stemplots, histograms) and the 68–95–99.7 rule. You also learned that a Normal probability plot is a great way to determine whether the shape of a distribution is approximately Normal. The more linear the Normal probability plot, the more Normal the distribution of the data.

What Did You Learn?

Learning Objective	Section	Related Example on Page(s)	Relevant Chapter Review Exercise(s)
Find and interpret the percentile of an individual value within a distribution of data.	2.1	86	R2.1
Estimate percentiles and individual values using a cumulative relative frequency graph.	2.1	87, 88	R2.2
Find and interpret the standardized score (*z*-score) of an individual value within a distribution of data.	2.1	90, 91	R2.1
Describe the effect of adding, subtracting, multiplying by, or dividing by a constant on the shape, center, and spread of a distribution of data.	2.1	93, 94, 95	R2.3
Estimate the relative locations of the median and mean on a density curve.	2.2	Discussion on 106–107	R2.4
Use the 68–95–99.7 rule to estimate areas (proportions of values) in a Normal distribution.	2.2	111	R2.5
Use Table A or technology to find (i) the proportion of *z*-values in a specified interval, or (ii) a *z*-score from a percentile in the standard Normal distribution.	2.2	114, 115 Discussion on 116	R2.6
Use Table A or technology to find (i) the proportion of values in a specified interval, or (ii) the value that corresponds to a given percentile in any Normal distribution.	2.2	118, 119, 120	R2.7, R2.8, R2.9
Determine whether a distribution of data is approximately Normal from graphical and numerical evidence.	2.2	122, 123, 124	R2.10, R2.11

Chapter 2 Chapter Review Exercises

These exercises are designed to help you review the important ideas and methods of the chapter.

R2.1 Is Paul tall? According to the National Center for Health Statistics, the distribution of heights for 15-year-old males has a mean of 170 centimeters (cm) and a standard deviation of 7.5 cm. Paul is 15 years old and 179 cm tall.

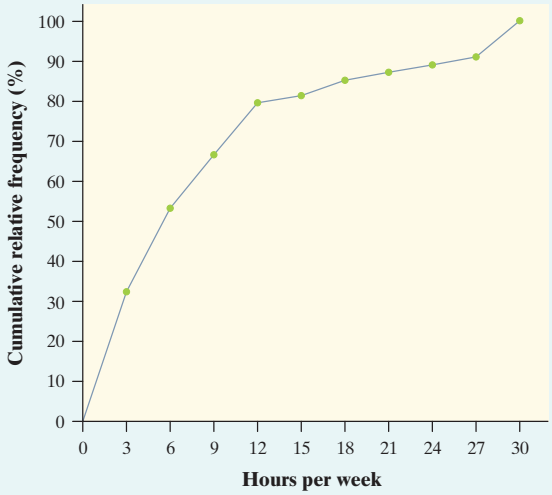
(a) Find the z-score corresponding to Paul's height. Explain what this value means.

(b) Paul's height puts him at the 85th percentile among 15-year-old males. Explain what this means to someone who knows no statistics.

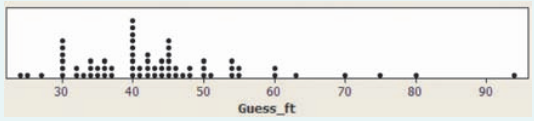
R2.2 Computer use Mrs. Causey asked her students how much time they had spent using a computer during the previous week. The following figure shows a cumulative relative frequency graph of her students' responses.

(a) At what percentile does a student who used her computer for 7 hours last week fall?

(b) Estimate the interquartile range (*IQR*) from the graph. Show your work.



R2.3 Aussie, Aussie, Aussie A group of Australian students were asked to estimate the width of their classroom in feet. Use the dotplot and summary statistics below to answer the following questions.



Variable	n	Mean	Stdev	Minimum	Q_1	Median	Q_3	Maximum
Guess ft	66	43.70	12.50	24.00	35.50	42.00	48.00	94.00

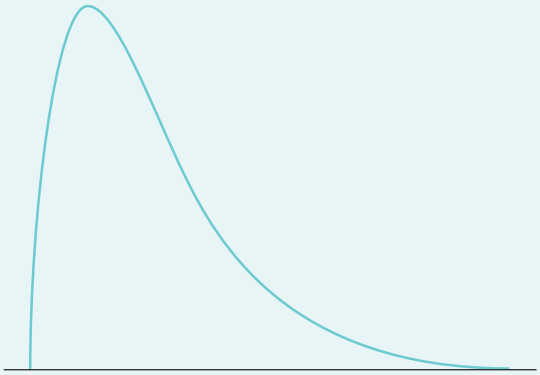
(a) Suppose we converted each student's guess from feet to meters (3.28 ft = 1 m). How would the shape of the distribution be affected? Find the mean, median, standard deviation, and *IQR* for the transformed data.

(b) The actual width of the room was 42.6 feet. Suppose we calculated the error in each student's guess as follows: guess − 42.6. Find the mean and standard deviation of the errors. Justify your answers.

R2.4 What the mean means The figure below is a density curve. Trace the curve onto your paper.

(a) Mark the approximate location of the median. Explain your choice of location.

(b) Mark the approximate location of the mean. Explain your choice of location.



R2.5 Horse pregnancies Bigger animals tend to carry their young longer before birth. The length of horse pregnancies from conception to birth varies according to a roughly Normal distribution with mean 336 days and standard deviation 3 days. Use the 68–95–99.7 rule to answer the following questions.

(a) Almost all (99.7%) horse pregnancies fall in what interval of lengths?

(b) What percent of horse pregnancies are longer than 339 days? Show your work.

R2.6 Standard Normal distribution Use Table A (or technology) to find each of the following for a standard Normal distribution. In each case, sketch a standard Normal curve and shade the area of interest.

(a) The proportion of observations with $-2.25 < z < 1.77$

(b) The number z such that 35% of all observations are greater than z

R2.7 **Low-birth-weight babies** Researchers in Norway analyzed data on the birth weights of 400,000 newborns over a 6-year period. The distribution of birth weights is Normal with a mean of 3668 grams and a standard deviation of 511 grams.[16] Babies that weigh less than 2500 grams at birth are classified as "low birth weight."

(a) What percent of babies will be identified as low birth weight? Show your work.

(b) Find the quartiles of the birth weight distribution. Show your work.

R2.8 **Ketchup** A fast-food restaurant has just installed a new automatic ketchup dispenser for use in preparing its burgers. The amount of ketchup dispensed by the machine follows a Normal distribution with mean 1.05 ounces and standard deviation 0.08 ounce.

(a) If the restaurant's goal is to put between 1 and 1.2 ounces of ketchup on each burger, what percent of the time will this happen? Show your work.

(b) Suppose that the manager adjusts the machine's settings so that the mean amount of ketchup dispensed is 1.1 ounces. How much does the machine's standard deviation have to be reduced to ensure that at least 99% of the restaurant's burgers have between 1 and 1.2 ounces of ketchup on them?

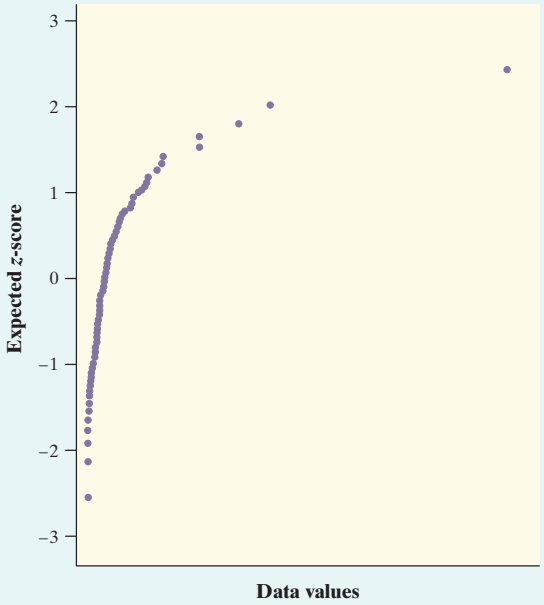
R2.9 **Grading managers** Many companies "grade on a bell curve" to compare the performance of their managers and professional workers. This forces the use of some low performance ratings, so that not all workers are listed as "above average." Ford Motor Company's "performance management process" for a time assigned 10% A grades, 80% B grades, and 10% C grades to the company's 18,000 managers. Suppose that Ford's performance scores really are Normally distributed. This year, managers with scores less than 25 received C's, and those with scores above 475 received A's. What are the mean and standard deviation of the scores? Show your work.

R2.10 **Fruit fly thorax lengths** Here are the lengths in millimeters of the thorax for 49 male fruit flies:[17]

0.64	0.64	0.64	0.68	0.68	0.68	0.72	0.72	0.72	0.72	0.74	0.76	0.76
0.76	0.76	0.76	0.76	0.76	0.76	0.78	0.80	0.80	0.80	0.80	0.80	0.82
0.82	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.88	0.88
0.88	0.88	0.88	0.88	0.88	0.88	0.92	0.92	0.92	0.94			

Are these data approximately Normally distributed? Give appropriate graphical and numerical evidence to support your answer.

R2.11 **Assessing Normality** A Normal probability plot of a set of data is shown here. Would you say that these measurements are approximately Normally distributed? Why or why not?



Chapter 2 AP® Statistics Practice Test

Section I: Multiple Choice *Select the best answer for each question.*

T2.1 Many professional schools require applicants to take a standardized test. Suppose that 1000 students take such a test. Several weeks after the test, Pete receives his score report: he got a 63, which placed him at the 73rd percentile. This means that

(a) Pete's score was below the median.

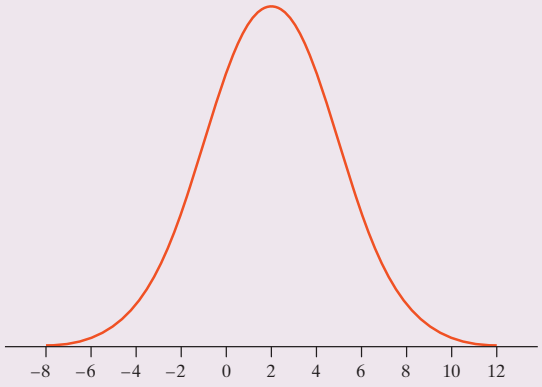
(b) Pete did worse than about 63% of the test takers.

(c) Pete did worse than about 73% of the test takers.

(d) Pete did better than about 63% of the test takers.

(e) Pete did better than about 73% of the test takers.

T2.2 For the Normal distribution shown, the standard deviation is closest to



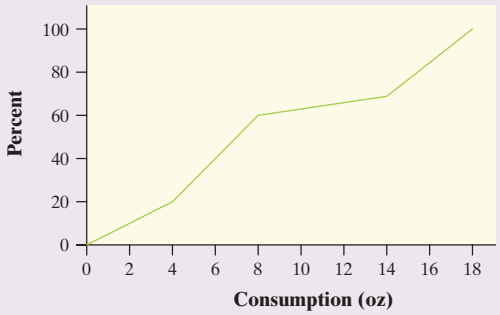
(a) 0 **(b)** 1 **(c)** 2 **(d)** 3 **(e)** 5

T2.3 Rainwater was collected in water collectors at 30 different sites near an industrial complex, and the amount of acidity (pH level) was measured. The mean and standard deviation of the values are 4.60 and 1.10, respectively. When the pH meter was recalibrated back at the laboratory, it was found to be in error. The error can be corrected by adding 0.1 pH units to all of the values and then multiplying the result by 1.2. The mean and standard deviation of the corrected pH measurements are

(a) 5.64, 1.44 **(c)** 5.40, 1.44 **(e)** 5.64, 1.20

(b) 5.64, 1.32 **(d)** 5.40, 1.32

T2.4 The figure shows a cumulative relative frequency graph of the number of ounces of alcohol consumed per week in a sample of 150 adults who report drinking alcohol occasionally. About what percent of these adults consume between 4 and 8 ounces per week?



(a) 20% **(b)** 40% **(c)** 50% **(d)** 60% **(e)** 80%

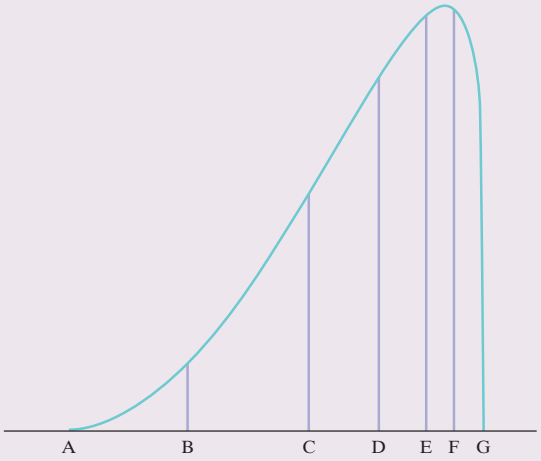
T2.5 The average yearly snowfall in Chillyville is Normally distributed with a mean of 55 inches. If the snowfall in Chillyville exceeds 60 inches in 15% of the years, what is the standard deviation?

(a) 4.83 inches **(d)** 8.93 inches

(b) 5.18 inches **(e)** The standard deviation cannot be computed from the given information.

(c) 6.04 inches

T2.6 The figure shown is the density curve of a distribution. Seven values are marked on the density curve. Which of the following statements is true?



(a) The mean of the distribution is E.

(b) The area between B and F is 0.50.

(c) The median of the distribution is C.

(d) The 3rd quartile of the distribution is D.

(e) The area between A and G is 1.

T2.7 If the heights of a population of men follow a Normal distribution, and 99.7% have heights between 5′0″ and 7′0″, what is your estimate of the standard deviation of the heights in this population?

(a) 1″ **(b)** 3″ **(c)** 4″ **(d)** 6″ **(e)** 12″

T2.8 Which of the following is *not* correct about a standard Normal distribution?

(a) The proportion of scores that satisfy $0 < z < 1.5$ is 0.4332.

(b) The proportion of scores that satisfy $z < -1.0$ is 0.1587.

(c) The proportion of scores that satisfy $z > 2.0$ is 0.0228.

(d) The proportion of scores that satisfy $z < 1.5$ is 0.9332.

(e) The proportion of scores that satisfy $z > -3.0$ is 0.9938.

Questions T2.9 and T2.10 refer to the following setting. Until the scale was changed in 1995, SAT scores were based on a scale set many years ago. For Math scores, the mean under the old scale in the 1990s was 470 and the standard deviation was 110. In 2009, the mean was 515 and the standard deviation was 116.

T2.9 What is the standardized score (z-score) for a student who scored 500 on the old SAT scale?

(a) −30 **(b)** −0.27 **(c)** −0.13 **(d)** 0.13 **(e)** 0.27

T2.10 Gina took the SAT in 1994 and scored 500. Her cousin Colleen took the SAT in 2013 and scored 530. Who did better on the exam, and how can you tell?

(a) Colleen—she scored 30 points higher than Gina.

(b) Colleen—her standardized score is higher than Gina's.

(c) Gina—her standardized score is higher than Colleen's.

(d) Gina—the standard deviation was bigger in 2013.

(e) The two cousins did equally well—their z-scores are the same.

Section II: Free Response *Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.*

T2.11 As part of the President's Challenge, students can attempt to earn the Presidential Physical Fitness Award or the National Physical Fitness Award by meeting qualifying standards in five events: curl-ups, shuttle run, sit and reach, one-mile run, and pull-ups. The qualifying standards are based on the 1985 School Population Fitness Survey. For the Presidential Award, the standard for each event is the 85th percentile of the results for a specific age group and gender among students who participated in the 1985 survey. For the National Award, the standard is the 50th percentile. To win either award, a student must meet the qualifying standard for all five events.

Jane, who is 9 years old, did 40 curl-ups in one minute. Matt, who is 12 years old, also did 40 curl-ups in one minute. The qualifying standard for the Presidential Award is 39 curl-ups for Jane and 50 curl-ups for Matt. For the National Award, the standards are 30 and 40, respectively.

(a) Compare Jane's and Matt's performances using percentiles. Explain in language simple enough for someone who knows little statistics to understand.

(b) Who has the higher standardized score (z-score), Jane or Matt? Justify your answer.

T2.12 The army reports that the distribution of head circumference among male soldiers is approximately Normal with mean 22.8 inches and standard deviation 1.1 inches.

(a) A male soldier whose head circumference is 23.9 inches would be at what percentile? Show your method clearly.

(b) The army's helmet supplier regularly stocks helmets that fit male soldiers with head circumferences between 20 and 26 inches. Anyone with a head circumference outside that interval requires a customized helmet order. What percent of male soldiers require custom helmets?

(c) Find the interquartile range for the distribution of head circumference among male soldiers.

T2.13 A study recorded the amount of oil recovered from the 64 wells in an oil field. Here are descriptive statistics for that set of data from Minitab.

Descriptive Statistics: Oilprod

Variable	n	Mean	Median	StDev	Min	Max	Q_1	Q_3
Oilprod	64	48.25	37.80	40.24	2.00	204.90	21.40	60.75

Does the amount of oil recovered from all wells in this field seem to follow a Normal distribution? Give appropriate statistical evidence to support your answer.

Chapter 2

Section 2.1

Answers to Check Your Understanding

page 89: **1.** c **2.** Her daughter weighs more than 87% of girls her age and she is taller than 67% of girls her age. **3.** About 65% of calls lasted less than 30 minutes, which means that about 35% of calls lasted 30 minutes or longer. **4.** $Q_1 = 13$ minutes, $Q_3 = 32$ minutes, and $IQR = 19$ minutes.

page 91: **1.** $z = -0.466$. Lynette's height is 0.466 standard deviations below the mean height of the class. **2.** $z = 1.63$. Brent's height is 1.63 standard deviations above the mean height of the class. **3.** $-0.85 = \dfrac{74 - 76}{\sigma}$, so $\sigma = 2.35$ inches.

page 97: **1.** Shape will not change. However, it will multiply the center (mean, median) and spread (range, *IQR*, standard deviation) by 2.54. **2.** Shape and spread will not change. It will, however, add 6 inches to the center (mean, median). **3.** Shape will not change. However, it will change the mean to 0 and the standard deviation to 1.

Answers to Odd-Numbered Section 2.1 Exercises

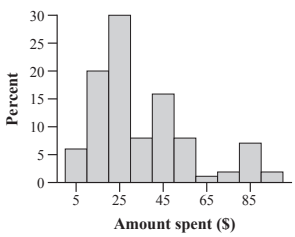
2.1 **(a)** She is at the 25th percentile, meaning that 25% of the girls had fewer pairs of shoes than she did. **(b)** He is at the 85th percentile, meaning that 85% of the boys had fewer pairs of shoes than he did. **(c)** The boy is more unusual because only 15% of the boys have as many or more than he has. The girl has a value that is closer to the center of the distribution.

2.3 A percentile only describes the relative location of a value in a distribution. Scoring at the 60th percentile means that Josh's score is better than 60% of the students taking this test. His correct percentage could be greater than 60% or less than 60%, depending on the difficulty of the test.

2.5 The girl weighs more than 48% of girls her age, but is taller than 78% of the girls her age.

2.7 **(a)** The student sent about 205 text messages in the 2-day period and sent more texts than about 78% of the students in the sample. **(b)** Locate 50% on the *y*-axis, read over to the points, and then go down to the *x*-axis. The median is approximately 115 text messages.

2.9 **(a)** $IQR \approx \$46 - \$19 = \$27$ **(b)** About the 26th percentile. **(c)** The histogram is below.



2.11 Eleanor. Her standardized score ($z = 1.8$) is higher than Gerald's ($z = 1.5$).

2.13 **(a)** Your bone density is far below average—about 1.5 times farther below average than a typical below-average density. **(b)** Solving $-1.45 = \dfrac{948 - 956}{\sigma}$ gives $\sigma = 5.52$ g/cm^2.

2.15 **(a)** He is at the 76th percentile, meaning his salary is higher than 76% of his teammates. **(b)** $z = 0.79$. Lidge's salary was 0.79 standard deviations above the mean salary.

2.17 Multiply each score by 4 and add 27.

2.19 **(a)** mean = 87.188 inches and median = 87.5 inches. **(b)** The standard deviation (3.20 inches) and *IQR* (3.25 inches) do not change because adding a constant to each value in a distribution does not change the spread.

2.21 **(a)** mean = 5.77 feet and median = 5.79 feet. **(b)** Standard deviation = 0.267 feet and *IQR* = 0.271 feet.

2.23 Mean $= \dfrac{9}{5}(25) + 32 = 77$°F and standard deviation $= \dfrac{9}{5}(2) = 3.6$°F.

2.25 c

2.27 c

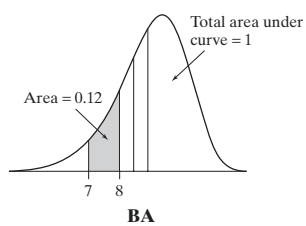
2.29 c

2.31 The distribution is skewed to the right with a center around 20 minutes and the range close to 90 minutes. The two largest values appear to be outliers.

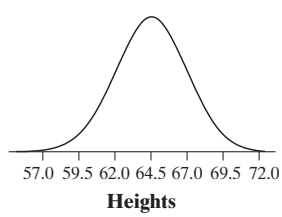
Section 2.2

Answers to Check Your Understanding

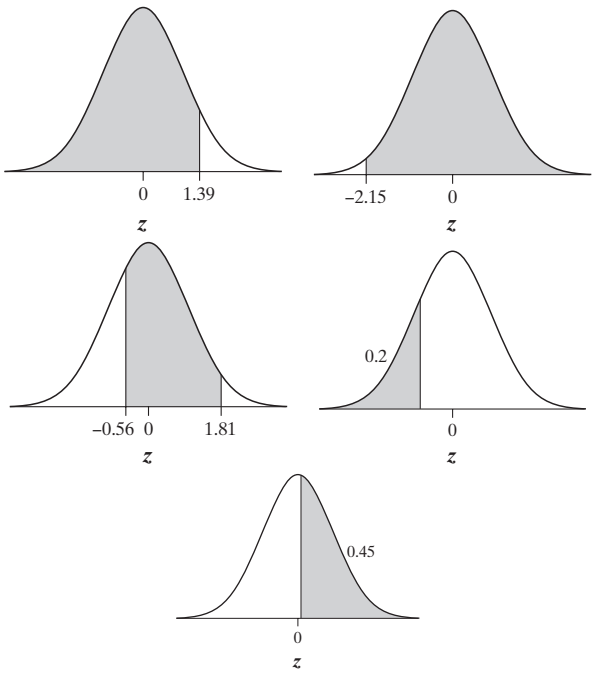
page 107: **1.** It is legitimate because it is positive everywhere and it has total area under the curve = 1. **2.** 12% **3.** Point A in the graph below is the approximate median. About half of the area is to the left of A and half of the area is to the right of A. **4.** Point B in the graph below is the approximate mean (balance point). The mean is less than the median in this case because the distribution is skewed to the left.



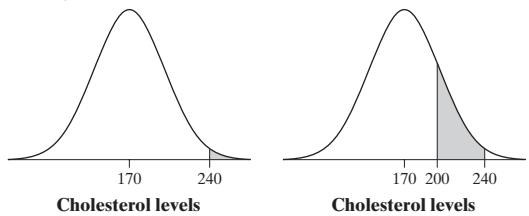
page 112: **1.** The graph is given below. **2.** Approximately $\dfrac{100\% - 68\%}{2} = 16\%$. **3.** Approximately $\dfrac{100\% - 68\%}{2} = 16\%$ have heights below 62 inches and approximately $\dfrac{100\% - 99.7\%}{2} = 0.15\%$ of young women have heights above 72 inches, so the remaining 83.85% have heights between 62 and 72 inches.

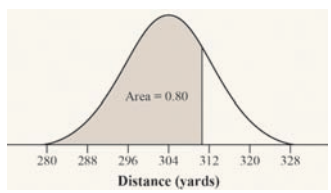


page 116: (All graphs are shown on the following page.) **1.** The proportion is 0.9177. **2.** The proportion is 0.9842. **3.** The proportion is 0.9649 − 0.2877 = 0.6772. **4.** The *z*-score for the 20th percentile is $z = -0.84$. **5.** 45% of the observations are greater than $z = 0.13$.



page 121: **1.** For 14-year-old boys, the amount of cholesterol follows a N(170, 30) distribution and we want to find the percent of boys with cholesterol of more than 240 (see graph below). $z = \dfrac{240 - 170}{30} = 2.33$. From Table A, the proportion of z-scores above 2.33 is $1 - 0.9901 = 0.0099$. *Using technology:* `normalcdf(lower:240,upper:1000,μ:170,σ:30)` = 0.0098. About 1% of 14-year-old boys have cholesterol above 240 mg/dl. **2.** For 14-year-old boys, the amount of cholesterol follows a N(170, 30) distribution and we want to find the percent of boys with cholesterol between 200 and 240 (see graph below). $z = \dfrac{200 - 170}{30} = 1$ and $z = \dfrac{240 - 170}{30} = 2.33$. From Table A, the proportion of z-scores between 1 and 2.33 is $0.9901 - 0.8413 = 0.1488$. *Using technology:* `normalcdf(lower:200,upper:240,μ:170,σ:30)` = 0.1488. About 15% of 14-year-old boys have cholesterol between 200 and 240 mg/dl. **3.** For Tiger Woods, the distance his drives travel follows an N (304, 8) distribution and the 80th percentile is the boundary value x with 80% of the distribution to its left (see graph below). A z-score of 0.84 gives the area closest to 0.80 (0.7995). Solving $0.84 = \dfrac{x - 304}{8}$ gives $x = 310.7$. *Using technology:* `invNorm(area:0.8,μ:304,σ:8)` = 310.7. The 80th percentile of Tiger Woods's drive lengths is about 310.7 yards.





Answers to Odd-Numbered Section 2.2 Exercises

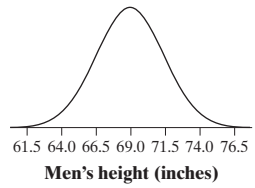
2.33 Sketches will vary, but here is one example:

2.35 **(a)** It is on or above the horizontal axis everywhere, and the area beneath the curve is $\frac{1}{3} \times 3 = 1$. **(b)** $\frac{1}{3} \times 1 = \frac{1}{3}$. **(c)** Because $1.1 - 0.8 = 0.3$, the proportion is $\frac{1}{3} \times 0.3 = 0.1$.

2.37 Both are 1.5.

2.39 **(a)** Mean is C, median is B. **(b)** Mean is B, median is B.

2.41 The graph is shown below.



2.43 **(a)** Between $69 - 2(2.5) = 64$ and $69 + 2(2.5) = 74$ inches. **(b)** About $\dfrac{100\% - 95\%}{2} = 2.5\%$. **(c)** About $\dfrac{100\% - 68\%}{2} = 16\%$ of men are shorter than 66.5 inches and $\dfrac{100\% - 95\%}{2} = 2.5\%$ are shorter than 64 inches, so approximately $16\% - 2.5\% = 13.5\%$ of men have heights between 64 inches and 66.5 inches. **(d)** Because $\dfrac{100\% - 68\%}{2} = 16\%$ of the area is to the right of 71.5, 71.5 is at the 84th percentile.

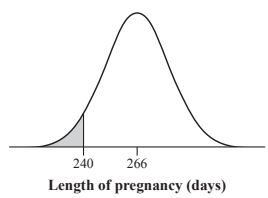
2.45 Taller curve: standard deviation ≈ 0.2. Shorter curve: standard deviation ≈ 0.5.

2.47 **(a)** 0.9978. **(b)** $1 - 0.9978 = 0.0022$ **(c)** $1 - 0.0485 = 0.9515$ **(d)** $0.9978 - 0.0485 = 0.9493$

2.49 **(a)** $0.9505 - 0.0918 = 0.8587$ **(b)** $0.9633 - 0.6915 = 0.2718$

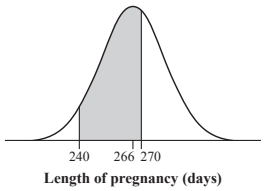
2.51 **(a)** $z = -1.28$ **(b)** $z = 0.41$

2.53 **(a)** The length of pregnancies follows a N(266, 16) distribution and we want the proportion of pregnancies that last less than 240 days (see graph below). $z = \dfrac{240 - 266}{16} = -1.63$. From Table A, the proportion of z-scores less than -1.63 is 0.0516. *Using technology:* `normalcdf(lower:-1000,upper:240,μ:266,σ:16)` = 0.0521. About 5% of pregnancies last less than 240 days, so 240 days is at the 5th percentile of pregnancy lengths.

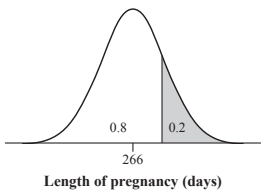


(b) The length of pregnancies follows a N(266, 16) distribution and we want the proportion of pregnancies that last between 240 and 270 days (see the following graph). $z = \dfrac{240 - 266}{16} = -1.63$ and $z = \dfrac{270 - 266}{16} = 0.25$. From Table A, the proportion of z-scores between -1.63 and 0.25 is $0.5987 - 0.0516 = 0.5471$. *Using technology:* `normalcdf(lower:240,upper:270,`

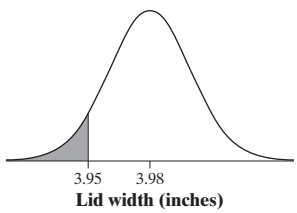
μ:266,σ:16) = 0.5466. About 55% of pregnancies last between 240 and 270 days.



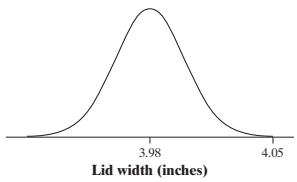
(c) The length of pregnancies follows a $N(266, 16)$ distribution and we are looking for the boundary value x that has an area of 0.20 to the right and 0.80 to the left (see graph below). A z-score of 0.84 gives the area closest to 0.80 (0.7995). Solving $0.84 = \frac{x - 266}{16}$ gives $x = 279.44$. *Using technology:* `invNorm(area:0.8,μ:266, σ:16)` = 279.47. The longest 20% of pregnancies last longer than 279.47 days.



2.55 (a) For large lids, the diameter follows a $N(3.98, 0.02)$ distribution and we want to find the percent of lids that have diameters less than 3.95 (see graph below). $z = \frac{3.95 - 3.98}{0.02} = -1.5$. From Table A, the proportion of z-scores below -1.5 is 0.0668. *Using technology:* `normalcdf(lower:-1000,upper:3.95, μ:3.98,σ:0.02)` = 0.0668. About 7% of the large lids are too small to fit.

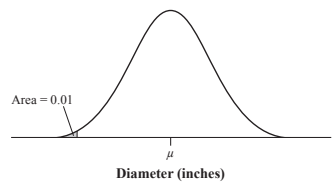


(b) For large lids, the diameter follows a $N(3.98, 0.02)$ distribution and we want to find the percent of lids that have diameters greater than 4.05 (see graph below). $z = \frac{4.05 - 3.98}{0.02} = 3.5$. From Table A, the proportion of z-scores above 3.50 is approximately 0. *Using technology:* `normalcdf(lower:4.05,upper:1000,μ:3.98, σ:0.02)` = 0.0002. Approximately 0% of the large lids are too big to fit.

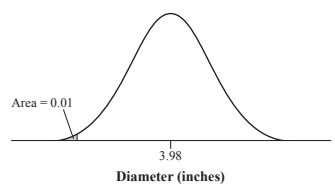


(c) Make a larger proportion of lids too small. If lids are too small, customers will just try another lid. But if lids are too large, the customer may not notice and then spill the drink.

2.57 (a) For large lids, the diameter follows a $N(\mu, 0.02)$ distribution and we want to find the value of μ that will result in only 1% of lids that are too small to fit (see graph below). A z-score of -2.33 gives the value closest to 0.01 (0.0099). Solving $-2.33 = \frac{3.95 - \mu}{0.02}$ gives $\mu = 4.00$. *Using technology:* `invNorm(area:0.01,μ:0,σ:1)` gives $z = -2.326$. Solving $-2.326 = \frac{3.95 - \mu}{0.02}$ gives $\mu = 4.00$. The manufacturer should set the mean diameter to approximately $\mu = 4.00$ to ensure that only 1% of lids are too small. (b) For large lids, the diameter follows a $N(3.98, \sigma)$ distribution and we want to find the value of σ that will result in only 1% of lids that are too small to fit (see graph below). A z-score of -2.33 gives the value closest to 0.01 (0.0099). Solving $-2.33 = \frac{3.95 - 3.98}{\sigma}$ gives $\sigma = 0.013$. *Using technology:* `invNorm(area:0.01,μ:0,σ:1)` gives $z = -2.326$. Solving $-2.326 = \frac{3.95 - 3.98}{\sigma}$ gives $\sigma = 0.013$. A standard deviation of at most 0.013 will result in only 1% of lids that are too small to fit.



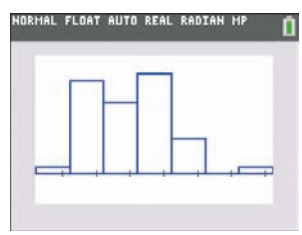
(c) Reduce the standard deviation. This will reduce the number of lids that are too small and the number of lids that are too big. If we make the mean a little larger as in part (a), we will reduce the number of lids that are too small, but we will increase the number of lids that are too big.



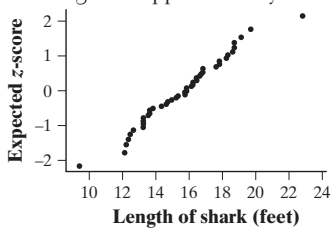
2.59 (a) $z = -1.28$ and $z = 1.28$ (b) Solving $-1.28 = \frac{x - 64.5}{2.5}$ gives $x = 61.3$ inches and solving $1.28 = \frac{x - 64.5}{2.5}$ gives $x = 67.7$ inches.

2.61 Solving $1.04 = \frac{60 - \mu}{\sigma}$ and $1.88 = \frac{75 - \mu}{\sigma}$ gives $\mu = 41.43$ minutes and $\sigma = 17.86$ minutes.

2.63 (a) A histogram is given below. The distribution of shark lengths is roughly symmetric and somewhat bell-shaped, with a mean of 15.586 feet and a standard deviation of 2.55 feet. (b) 30/44 = 68.2% , 42/44 = 95.5%, and 44/44 = 100%. These are very close to the 68–95–99.7 rule.



(c) A Normal probability plot is given below. Except for one small shark and one large shark, the plot is fairly linear, indicating that the distribution of shark lengths is approximately Normal.



(d) All indicate that shark lengths are approximately Normal.

2.65 The distribution is close to Normal because the plot is nearly linear. There is a small "wiggle" between 120 and 130, with several values a little larger than would be expected in a Normal distribution. Also, the smallest value and the two largest values are a little farther from the mean than would be expected in a Normal distribution.

2.67 No. If it was Normal, then the minimum value should be around 2 or 3 standard deviations below the mean. However, the actual minimum has a z-score of just $z = -1.09$. Also, if the distribution was Normal, the minimum and maximum should be about the same distance from the mean. However, the maximum is much farther from the mean (20,209) than the minimum (8741).

2.69 b

2.71 b

2.73 a

2.75 For both kinds of cars, we see that the highway mileage is greater than the city mileage. The two-seater cars have a more variable distribution, both on the highway and in the city. Also the mileage values are slightly lower for the two-seater cars than for the minicompact cars, both on the highway and in the city, with a greater difference on the highway. All four distributions are roughly symmetric.

Answers to Chapter 2 Review Exercises

R2.1 (a) $z = 1.20$. Paul's height is 1.20 standard deviations above the average male height for his age. (b) 85% of boys Paul's age are shorter than Paul.

R2.2 (a) 58th percentile (b) $IQR = 11 - 2.5 = 8.5$ hours per week.

R2.3 (a) The shape of the distribution would not change.

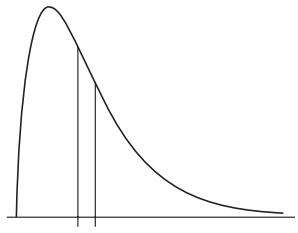
$$\text{Mean} = \frac{43.7}{3.28} = 13.32 \text{ meters, median} = \frac{42}{3.28} = 12.80 \text{ meters,}$$

$$\text{standard deviation} = \frac{12.5}{3.28} = 3.81 \text{ meters,}$$

$$IQR = \frac{12.5}{3.28} = 3.81 \text{ meters.}$$

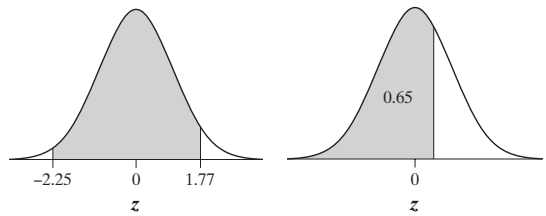
(b) Mean = 43.7 − 42.6 = 1.1 feet; standard deviation = 12.5 feet, because subtracting a constant from each observation does not change the spread.

R2.4 (a) The median (line A in the graph below) should be slightly to the right of the main peak, with half of the area to the left and half to the right. (b) The mean (line B in the graph below) should be slightly to the right of the line for the median at the balancing point.

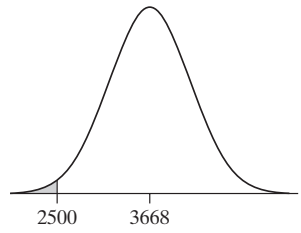


R2.5 (a) Between 336 − 3(3) = 327 days and 336 + 3(3) = 345 days. (b) About $\frac{100\% - 68\%}{2} = 16\%$.

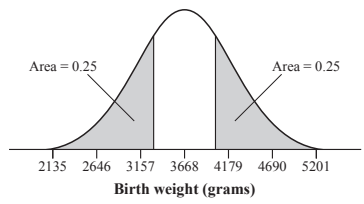
R2.6 (a) 0.9616 − 0.0122 = 0.9494 (b) If 35% of all values are greater than a particular z-value, then 65% are lower. A z-score of 0.39 gives the value closest to 0.65 (0.6517). *Using technology:* `invNorm(area:0.65,μ:0,σ:1)` gives $z = 0.385$.



R2.7 (a) Birth weights follow a N(3668, 511) distribution and we want to find the percent of babies with weights less than 2500 grams (see graph below). $z = \frac{2500 - 3668}{511} = -2.29$. From Table A, the proportion of z-scores below −2.29 is 0.0110. *Using technology:* `normalcdf(lower:-1000,upper:2500,μ:3668,σ:511)` = 0.0111. About 1% of babies will be identified as low birth weight.

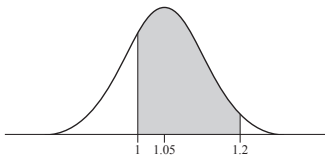


(b) Birth weights follow a N(3668, 511) distribution. The 1st quartile is the boundary value with 25% of the area to its left. The 3rd quartile is the boundary value with 75% of the area to its left (see graph below). A z-score of −0.67 gives the value closest to 0.25 (0.2514). Solving $-0.67 = \frac{x - 3668}{511}$ gives $Q_1 = 3325.63$. A z-score of 0.67 gives the value closest to 0.75 (0.7486). Solving $0.67 = \frac{x - 3668}{511}$ gives $Q_3 = 4010.37$. *Using technology:* `invNorm(area:0.25,μ:3668,σ:511)` gives $Q_1 = 3323.34$ and `invNorm(area:0.75,μ:3668,σ:511)` gives $Q_3 = 4012.66$. The quartiles are $Q_1 = 3323.34$ grams and $Q_3 = 4012.66$ grams.

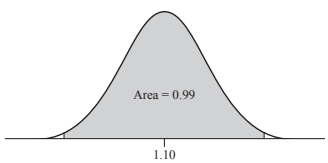


R2.8 (a) The amount of ketchup dispensed follows a N(1.05, 0.08) distribution and we want to find the percent of times that the amount of ketchup dispensed will be between 1 and 1.2 ounces (see

graph below). $z = \frac{1.2 - 1.05}{0.08} = 1.88$ and $z = \frac{1 - 1.05}{0.08} = -0.63$. From Table A, the proportion of z-scores between -0.63 and 1.88 is $0.9699 - 0.2643 = 0.7056$. *Using technology:* `normalcdf(lower:1,upper:1.2,μ:1.05,σ:0.08)` $= 0.7036$. About 70% of the time the dispenser will put between 1 and 1.2 ounces of ketchup on a burger.

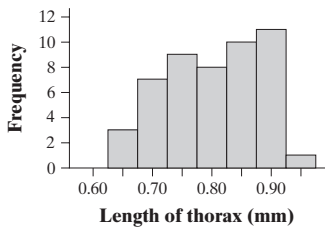


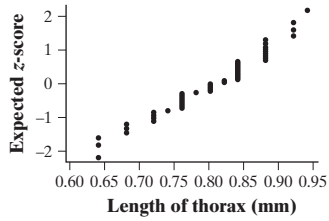
(b) The amount of ketchup dispensed follows a $N(1.1, \sigma)$ distribution and we want to find the value of σ that will result in at least 99% of burgers getting between 1 and 1.2 ounces of ketchup (see graph below). Because the mean of 1.1 is in the middle of the interval from 1 to 1.2, we are looking for the middle 99% of the distribution. This leaves 0.5% in each tail. A z-score of -2.58 gives the value closest to 0.005 (0.0049). Solving $-2.58 = \frac{1 - 1.1}{\sigma}$ gives $\sigma = 0.039$. *Using technology:* `invNorm(area:0.005,μ:0,σ:1)` gives $z = -2.576$. Solving $-2.576 = \frac{1 - 1.1}{\sigma}$ gives $\sigma = 0.039$. A standard deviation of at most 0.039 ounces will result in at least 99% of burgers getting between 1 and 1.2 ounces of ketchup.



R2.9 If the distribution is Normal, the 10th and 90th percentiles must be equal distances above and below the mean. Thus, the mean is $\frac{25 + 475}{2} = 250$ points. The 10th percentile in a standard Normal distribution is $z = -1.28$. Solving $-1.28 = \frac{25 - 250}{\sigma}$, we get $\sigma = 175.8$. *Using technology:* `invNorm(area:0.10,μ:0,σ:1)` gives $z = -1.282$, so $-1.282 = \frac{25 - 250}{\sigma}$ and $\sigma = 175.5$.

R2.10 A histogram and Normal probability plot are given below. The histogram is roughly symmetric but not very bell-shaped. The Normal probability plot, however, is roughly linear. For these data, $\bar{x} = 0.8004$ and $s_x = 0.0782$. Although the percentage within 1 standard deviation of the mean (55.1%) is less than expected (68%), the percentage within 2 (93.9%) and 3 standard deviations (100%) match the 68–95–99.7 rule quite well. It is reasonable to say that these data are approximately Normally distributed.





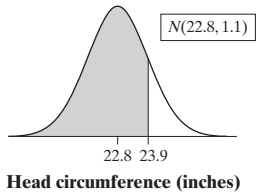
R2.11 The steep, nearly vertical portion at the bottom and the clear bend to the right indicate that the distribution of the data is right-skewed with several outliers and not approximately Normally distributed.

Answers to Chapter 2 AP® Statistics Practice Test

T2.1 e
T2.2 d
T2.3 b
T2.4 b
T2.5 a
T2.6 e
T2.7 c
T2.8 e
T2.9 e
T2.10 c

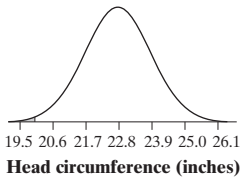
T2.11 **(a)** Jane's performance was better. Because her performance (40) exceeded the standard for the Presidential award (39), she performed above the 85th percentile. Matt's performance (40) met the standard for the National award (40), meaning he performed at the 50th percentile. **(b)** Because Jane's score has a higher percentile than Matt's score, she is farther to the right in her distribution than Matt is in his. Therefore, Jane's standardized score will likely be greater than Matt's.

T2.12 **(a)** For male soldiers, head circumference follows a $N(22.8, 1.1)$ distribution and we want to find the percent of soldiers with head circumference less than 23.9 inches (see graph below). $z = \frac{23.9 - 22.8}{1.1} = 1$. From Table A, the proportion of z-scores below 1 is 0.8413. *Using technology:* `normalcdf(lower:-1000,upper:23.9,μ:22.8,σ:1.1)` $= 0.8413$. About 84% of soldiers have head circumferences less than 23.9 inches. Thus, 23.9 inches is at the 84th percentile.

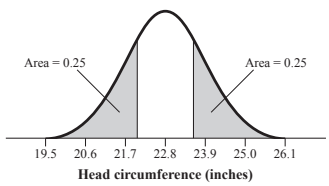


(b) For male soldiers, head circumference follows a $N(22.8, 1.1)$ distribution and we want to find the percent of soldiers with head circumferences less than 20 inches or greater than 26 inches (see graph below). $z = \frac{20 - 22.8}{1.1} = -2.55$ and $z = \frac{26 - 22.8}{1.1} = 2.91$. From Table A, the proportion of z-scores below $z = -2.55$ is 0.0054 and the proportion of z-scores above 2.91 is $1 - 0.9982 = 0.0018$, for a total of $0.0054 + 0.0018 = 0.0072$. *Using technology:* $1 -$ `normalcdf(lower:20,upper:26,μ:22.8,σ:1.1)` $= 1 - 0.9927 = 0.0073$. A little less than 1% of soldiers have head

circumferences less than 20 inches or greater than 26 inches and require custom helmets.



(c) For male soldiers, head circumference follows a N(22.8, 1.1) distribution. The 1st quartile is the boundary value with 25% of the area to its left. The 3rd quartile is the boundary value with 75% of the area to its left (see graph below). A z-score of -0.67 gives the value closest to 0.25 (0.2514). Solving $-0.67 = \dfrac{x - 22.8}{1.1}$ gives $Q_1 = 22.063$. A z-score of 0.67 gives the value closest to 0.75 (0.7486). Solving $0.67 = \dfrac{x - 22.8}{1.1}$ gives $Q_3 = 23.537$. *Using technology:* `invNorm(area:0.25,μ:22.8,σ:1.1)` gives $Q_1 = 22.058$ and `invNorm(area:0.75,μ:22.8,σ:1.1)` gives $Q_3 = 23.542$. Thus, $IQR = 23.542 - 22.058 = 1.484$ inches.



T2.13 No. First, there is a large difference between the mean and the median. In a Normal distribution, the mean and median are the same, but in this distribution the mean is 48.25 and the median is 37.80. Second, the distance between the minimum and the median is 35.80 but the distance between the median and the maximum is 167.10. In a Normal distribution, these distances should be about the same. Both of these facts suggest that the distribution is skewed to the right.

Chapter 3

Section 3.1

Answers to Check Your Understanding

page 144: **1.** Explanatory: number of cans of beer. Response: blood alcohol level. **2.** Explanatory: amount of debt and income. Response: stress caused by college debt.

page 149: **1.** Positive. The longer the duration of the eruption, the longer we should expect to wait between eruptions because long eruptions use more energy and it will take longer to build up the energy needed to erupt again. **2.** Roughly linear with two clusters. The clusters indicate that, in general, there are two types of eruptions—shorter eruptions that last around 2 minutes and longer eruptions that last around 4.5 minutes. **3.** Fairly strong. The points don't deviate much from the linear form. **4.** There are a few possible outliers around the clusters. However, there aren't many and potential outliers are not very distant from the main clusters of points. **5.** How long the previous eruption was.

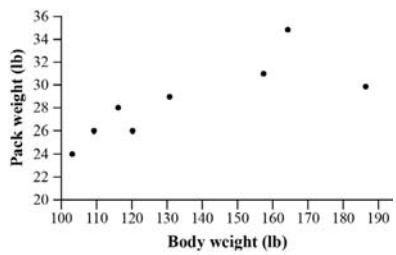
page 153: **(a)** $r \approx 0.9$. This indicates that there is a strong, positive linear relationship between the number of boats registered in Florida and the number of manatees killed. **(b)** $r \approx 0.5$. This indicates that there is a moderate, positive linear relationship between the number of named storms predicted and the actual number of named storms. **(c)** $r \approx 0.3$. This indicates that there is a weak, positive linear relationship between the healing rate of the two front limbs of the newts. **(d)** $r \approx -0.1$. This indicates that there is a weak, negative linear relationship between last year's percent return and this year's percent return in the stock market.

Answers to Odd-Numbered Section 3.1 Exercises

3.1 Explanatory: water temperature (quantitative). Response: weight change (quantitative).

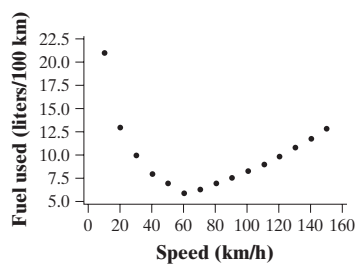
3.3 **(a)** Positive. Students with higher IQs tend to have higher GPAs and vice versa because both IQ and GPA are related to mental ability. **(b)** Roughly linear, because a line through the scatterplot of points would provide a good summary. Moderately strong, because most of the points would be close to the line. **(c)** IQ $\approx$ 103 and GPA $\approx$ 0.4.

3.5 A scatterplot is shown below.



3.7 **(a)** There is a positive association between backpack weight and body weight. For students under 140 pounds, there seems to be a linear pattern in the graph. However, for students above 140 pounds, the association begins to curve. Because the points vary somewhat from the linear pattern, the relationship is only moderately strong. **(b)** The hiker with body weight 187 pounds and pack weight 30 pounds. This hiker makes the form appear to be nonlinear for weights above 140 pounds. Without this hiker, the association would look very linear for all body weights.

3.9 **(a)** A scatterplot is shown below. **(b)** The relationship is curved. Large amounts of fuel were used for low and high values of speed and smaller amounts of fuel were used for moderate speeds. This makes sense because the best fuel efficiency is obtained by driving at moderate speeds. **(c)** Both directions are present in the scatterplot. The association is negative for lower speeds and positive for higher speeds. **(d)** The relationship is very strong, with little deviation from a curve that can be drawn through the points.



3.11 **(a)** Most of the southern states fall in the same pattern as the rest of the states. However, southern states typically have lower mean SAT math scores than other states with a similar percent of students taking the SAT. **(b)** West Virginia has a much lower mean